Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique

Cyril Grouin

École doctorale « Pierre Louis de santé publique à Paris »
ED 393 Épidémiologie et Sciences de l'Information Biomédicale
INSERM U872 Eq 20 *Ingénierie des Connaissances en Santé*, LIMSI–CNRS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Informatique Biomédicale

présentée par

Cyril GROUIN

pour obtenir le grade de Docteur de l'Université Pierre et Marie Curie

---

Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique

---

Soutenue le 26 juin 2013 devant le jury composé de :

Pr Stéfan J. DARMONI, PU-PH	CHU de Rouen	rapporteur
Pr Pascal STACCINI, PU-PH	CHU de Nice	rapporteur
M. Thierry ARTIÈRES, PU	UPMC, LIP6	examinateur
Pr Anita BURGUN, PU-PH	HEGP, INSERM U872 Eq 22	examinatrice
M^{me} Marie-Christine JAULENT, DR	INSERM U872 Eq 20	co-directrice
M. Pierre ZWEIGENBAUM, DR	CNRS, LIMSI	co-directeur

Informations administratives

Faculté de Biologie Pierre et Marie Curie
UFR 927 *Sciences de la Vie*
9, quai Saint-Bernard, 75005 Paris
http://www.biologie.upmc.fr/

École doctorale « Pierre Louis de santé publique à Paris »
ED 393 *Épidémiologie et Sciences de l'Information Biomédicale*
Centre de Recherche des Cordeliers
15, rue de l'École de Médecine, 75006 Paris
http://www.ed393.upmc.fr/

Laboratoires d'accueil

INSERM, UMR_S 872
Centre de Recherche des Cordeliers
Eq 20 *Ingénierie des Connaissances en Santé*
15, rue de l'École de Médecine, 75006 Paris
http://ics.upmc.fr/

UPR 3251, LIMSI–CNRS, groupe ILES
Information Langue Ecrite et Signée
Rue John von Neumann, 91400 Orsay
http://www.limsi.fr/

Début de l'hiver
Thèse à présent rédigée
Tombent les flocons

« Cogite roulait des yeux. Il trouvait toujours ses arguments imparables quand il les échafaudait dans sa tête. Il lisait certains ouvrages anciens, passait une éternité à réfléchir, une petite théorie se formait toute seule sous son crâne en une rangée de petits cubes luisants, puis sitôt qu'il en parlait, il se heurtait à l'incrédulité des membres de la faculté et il fallait toujours, toujours, que l'un d'eux pose une saleté de question idiote à laquelle il ne savait pas répondre sur le moment. Comment progresser contre des esprits pareils ? »

Le dernier continent
TERRY PRATCHETT

Remerciements

Je remercie chaleureusement :

Pierre ZWEIGENBAUM, mon Directeur de thèse, pour son enthousiasme permanent, sa curiosité naturelle débordante, sa disponibilité et son aide patiente, et Marie-Christine JAULENT, ma co-Directrice de thèse, pour son accueil au sein de l'équipe n° 20 « *Ingénierie des Connaissances en Santé* » de l'UMR_S 872 de l'Inserm, Stéfan J. DARMONI et Pascal STACCINI qui m'ont fait l'honneur d'être rapporteurs de cette thèse, Thierry ARTIÈRES et Anita BURGUN pour l'intérêt qu'ils ont témoigné envers ce travail de thèse en acceptant d'être examinateurs,

Anne VILNAT qui m'a poussé à m'investir dans ce travail de thèse,

Louise DELÉGER pour son amitié, l'anonymisation qu'elle a réalisé d'une partie du corpus utilisé dans cette thèse, son aide sur la simulation de Monte Carlo,

Sophie ROSSET pour sa bonne humeur communicative, ses nombreuses idées à exploiter *(toujours adéquates mais loin d'être évidentes à mettre en place)*, et les multiples discussions tant sur les différents aspects de mon travail que sur les sujets plus ou moins importants de la vie quotidienne,

Les utilisatrices de l'outil Medina pour leurs retours et pistes d'amélioration (Natalia GRABAR, Anne-Do PHAM, Aurélie NÉVÉOL et Marion RICHARD), Olivier GALIBERT pour les outils d'évaluation et de calcul d'accords inter-annotateurs, Thomas LAVERGNE pour ses relectures, son outil d'apprentissage statistique Wapiti, et ses conseils d'utilisation de l'outil, Karën FORT et les membres du groupe de travail « *accords inter-annotateurs* » pour les séances de travail stimulantes et la découverte de l'article état-de-l'art sur les différents coefficients d'accord, Marianna APIDIANAKI pour la vérification de mes citations en grec du serment d'Hippocrate : ευχαριστώ !

Mes collègues de bureau, Béatrice ARNULPHY et Anne-Lyse MINARD, toujours souriantes et qui ont dû me supporter durant toutes ces années, Véronique MORICEAU et son projet autour du *Club Gourmet*, les amis et collègues du LIMSI qui ont inconsidérément accepté de participer aux campagnes d'évaluation i2b2 *(avec pour certains, des cas de récidive d'une année sur l'autre)* alors que les périodes de test tombent chaque année au mois d'août, ainsi que Gabriel ILLOUZ (et son maître à penser Załikowsky) pour ses remarques humoristiques, ses idées ingénieuses en matière de programmation, et le défi qu'il m'a lancé de placer une formule avec une double intégrale dans ce manuscrit *(mais quelle idée parfaitement grotesque !)*, défi à moitié rempli qui explique la formule de la page 113...

Ma famille.

Résumé

Résumé

Ce travail porte sur l'anonymisation automatique de comptes rendus cliniques. L'anonymisation consiste à masquer les informations personnelles présentes dans les documents tout en préservant les informations cliniques. Cette étape est obligatoire pour utiliser des documents cliniques en dehors du parcours de soins, qu'il s'agisse de publication de cas d'étude ou en recherche scientifique *(mise au point d'outils informatiques de traitement du contenu des dossiers, recherche de cas similaire, etc.)*.

Nous avons défini douze catégories d'informations à traiter : nominatives *(noms, prénoms, etc.)* et numériques *(âges, dates, codes postaux, etc.)*.

Deux approches ont été utilisées pour anonymiser les documents, l'une dite « symbolique », à base de connaissances d'expert formalisées par des expressions régulières et la projection de lexiques, l'autre par apprentissage statistique au moyen de CRF de chaîne linéaire. Plusieurs expériences ont été menées parmi lesquelles l'utilisation simple ou enchaînée de chacune des deux approches.

Nous obtenons nos meilleurs résultats (F-mesure globale=0,922) en enchaînant les deux méthodes avec rassemblement des noms et prénoms en une seule catégorie (pour cette catégorie : rappel=0,953 et F-mesure=0,931).

Ce travail de thèse s'accompagne de la production de plusieurs ressources : un guide d'annotation, un corpus de référence de 562 documents dont 100 annotés en double avec adjudication et calculs de taux d'accord inter-annotateurs (κ=0,807 avant fusion) et un corpus anonymisé de 17 000 comptes rendus cliniques.

Mots-clefs

Anonymisation, comptes rendus médicaux, guide d'annotation, méthodes symboliques, apprentissage statistique, traitement automatique des langues.

Clinical Records De-Identification: Performances and Limits of Rule-based and Machine-Learning based Approaches

Abstract

This work focuses on the automatic de-identification of clinical records. The de-identification consists in concealing personal information within documents while preserving clinical data. This task is mandatory so as to use clinical records outside of the patient care process, for case study publications or in scientific research *(producing automatic system to process the documents, similar cases search, etc.)*.

We defined 12 categories of information to de-identify: nominative data *(last names, first names, etc.)* and numerical data *(ages, dates, zip codes, etc.)*.

Two approaches have been used to de-identify the documents, an expert knowledge based method using regular expressions and lexical mapping, and a machine-learning process based upon CRF. Several experiments have been performed including the use of each approach separately or in combination.

We achieved our best results (overall F-measure=0.922) while combining both approaches and merging last names and first names categories into a single one (recall=0.953 and F-measure=0.931 on this category).

This work is combined with the production of several resources: a guidelines, a gold standard corpus composed of 562 documents among them 100 double annotated with adjudication and inter-annotator agreement computation (κ=0.807 before merging) and a de-identified corpus of 17,000 clinical records.

Keywords

De-identification, clinical records, guidelines, rule-based methods, machine-learning based approach, natural language processing.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

BILOU : *Begin, In, Last, Out, Unit*, schéma d'annotation utilisé pour représenter la base d'apprentissage des systèmes d'apprentissage statistique.

BIO : *Begin, In, Out*, schéma d'annotation utilisé pour représenter la base d'apprentissage des systèmes d'apprentissage statistique.

CCAM : Classification Commune des Actes Médicaux, nomenclature établie pour coder les différents actes médicaux et définir les honoraires de ces actes, http://www.ameli.fr/accueil-de-la-ccam/index.php.

CCHMC : *Cincinnati Children's Hospital Medical Center*, centre médical de l'hôpital pour enfants, Cincinnati, OH, http://www.cchmc.org/

CIM : Classification Internationale des Maladies, classification des maladies et traumatismes. Elle est utilisée en milieu hospitalier pour traiter de la morbidité hospitalière. Les États-Unis utilisent la 9^{e} révision de cette classification (CIM-9) alors que les autres États membres de l'OMS utilisent la CIM-10 depuis 1994. L'adoption de la CIM-10 par les États-Unis, initialement prévue pour le 1er octobre 2013, a été repoussée d'une année. La CIM-10 comprend vingt-et-un chapitres segmentés en sections, catégories, et sous-catégories sous lesquels sont classées les maladies. Un code unique à trois chiffres est associé à chaque maladie. Le manuel de codage est accessible à l'adresse suivante : http://www.health.fgov.be/internet2Prd/groups/public/@public/@dg1/@datamanagement/documents/ie2divers/994482_fr.pdf.

CISMeF : Catalogue et Index des Sites Médicaux de langue Française, http://www.chu-rouen.fr/cismef/

CMI : Certificat médical initial. Document à l'origine de l'hospitalisation et précisant les motifs d'hospitalisation.

CNIL : Commission Nationale Informatique et Libertés, http://www.cnil.fr/

CRF : *Conditional Random Fields*, champs aléatoires conditionnels, l'un des modèles log-linéaires utilisés par les systèmes d'apprentissage statistique, adapté pour l'étiquetage en séquences.

CRH : Compte rendu d'hospitalisation. Document produit à la fin du séjour hospitalier qui formalise l'ensemble des informations cliniques produites pendant le séjour et qui précise les modalités et traitements de sortie.

CRO : Compte rendu opératoire. Document produit pendant le séjour hospitalier à l'occasion de l'opération chirurgicale effectuée.

cTAKES : *clinical Text Analysis and Knowledge Extraction System*, système d'extraction d'information depuis des documents cliniques, produit par l'institut Mayo Clinic, Boston, MA, https://wiki.nci.nih.gov/display/VKC/cTAKES+2.5

CUI : *Concept Unique Identifier*, identifiant unique des concepts du metathesaurus de l'UMLS.

DCI : Dénomination Commune Internationale, norme de nommage des substances actives dans les noms de médicaments composée de segments clés lisibles et prononçables dans la plupart des langues.

DOI : *Digital Object Identifier*, système d'identification pérenne des objets numériques. Un identifiant est attribué de manière unique à chaque objet *(dont les articles scientifiques)* et peut être décodé en passant par l'interface http://dx.doi.org/.

Enamex : *Entity Named Expression*, classe des entités nommées regroupant les noms de personnes, de lieux et d'organisations dans les tâches de repérage d'entités nommées des conférences MUC.

GATE : *General Architecture for Text Engineering*, plateforme de traitement automatique des langues, Sheffield University, Sheffield, UK, http://gate.co.uk/

HIDE : *Health Information De-identification*, outil d'anonymisation automatique fondé sur le formalisme des CRF, http://www.mathcs.emory.edu/hide/

HIPAA : *Health Insurance Portability and Accountability Act*, loi américaine à l'origine de l'obligation d'anonymiser les documents cliniques. Cette loi a défini de dix-huit identifiants devant faire l'objet d'une anonymisation. Le paragraphe 164.514 de cette loi précise que le risque de réidentification doit être le plus faible possible, http://www.hhs.gov/ocr/privacy/

HMM : *Hidden Markov Model*, modèles de Markov cachés, modèle statistique décrivant une suite d'états et de transitions.

i2b2 : *Informatics for Integrating Biology & the Bedside*, institut de recherche américain organisant une campagne d'évaluation annuelle en informatique médicale, http://www.i2b2.org/ et https://www.i2b2.org/NLP/ pour le challenge.

Mallet : *MAchine Learning for LanguagE Toolkit*, outil d'apprentissage statistique pour le traitement automatique des langues reposant sur plusieurs formalismes *(arbres de décision, entropie maximale, réseaux bayésiens naïfs)*, University of Massachusetts, Amherst, MA, http://mallet.cs.umass.edu/

MAT : *MITRE Annotation Toolkit*, outil d'annotation de corpus utilisé dans l'outil MIST, The MITRE Corporation, Bedford, MA.

Medina : *MEDical INformation Anonymization*, outil d'anonymisation automatique produit dans le cadre de cette thèse.

MedLEE : *Medical Language Extraction and Encoding system*, Columbia University, New York, NY, http://www.cat.columbia.edu/medlee.htm

MEDLINE : base de données de la littérature mondiale du domaine biomédical avec accès aux articles, Bethesda, MD. L'accès à la base de données est possible via l'interface PubMed, http://www.ncbi.nlm.nih.gov/pubmed

MeSH : *Medical Subject Headings*, thesaurus de termes biomédicaux utilisé pour indexer les articles présents dans MEDLINE, Bethesda, MD, http://www.ncbi.nlm.nih.gov/mesh

MIMIC : *Multiparameter Intelligent Monitoring in Intensive Care*, projet mené par le MIT pour peupler une base de données d'informations biomédicales, Boston, MA, http://mimic.physionet.org/

MIST : *MITRE Identification Scrubber Toolkit*, outil d'anonymisation automatique fondé sur le formalisme des CRF, The MITRE Corporation, Bedford, MA, http://mist-deid.sourceforge.net/

MUC : *Message Understanding Conference*, conférences en fouille de textes, la sixième édition a vu l'apparition du concept d'« entités nommées ».

NLM : *National Library of Medicine*, organisme de recherche américain, Bethesda, MD, http://www.nlm.nih.gov

Numex : *Number Expression*, classe des expressions numériques *(montants et pourcentages)* dans les tâches de repérage d'entités nommées des conférences MUC.

OMS : Organisation Mondiale de la Santé, Genève, http://www.who.int/fr/

PHI : *Protected Health Information*, informations sensibles personnelles qu'il importe d'anonymiser dans un document.

REN : Repérage d'Entités Nommées, tâche consistant à repérer des éléments appartenant à des catégories prédéfinies *(noms de personnes, noms de lieux, noms d'organisations, dates, durées, montants, etc.)*.

SER : *Slot Error Rate*, mesure d'évaluation composite évaluant à la fois la catégorisation et les frontières d'une entité nommée.

SVM : *Support Vector Machines*, séparateurs à vaste marge, l'un des formalismes utilisés par les systèmes d'apprentissage statistique.

TAL : Traitement Automatique des Langues, ensemble des processus informatiques et des applications permettant de traiter les spécificités des langues naturelles.

Timex : *Time Expression*, classe des expressions temporelles *(dates, heures, durées et fréquences)* dans les tâches de repérage d'entités nommées des conférences MUC.

UIMA : *Unstructured Information Management Applications*, format d'architecture logicielle pour le traitement automatique des langues, http://uima.apache.org/

UMLS : *Unified Medical Language System*, ensemble d'outils *(Metathesaurus, Specialist Lexicon, Semantic Network)* pour le traitement de la langue biomédicale, Bethesda, MD, http://www.nlm.nih.gov/research/umls/

VHA : *Veterans Health Administration*, division santé du département américain des vétérans, Washington, DC, http://www.va.gov/health/default.asp

Introduction générale

Contexte historique

Au IV^e^ siècle avant notre ère, en Grèce, fut rédigé un serment jetant les bases de la déontologie médicale. Traditionnellement attribué à Hippocrate, médecin grec aujourd'hui considéré comme le « *Père de la médecine* », ce serment dresse une liste de devoirs et d'engagements vis à vis des maîtres de la médecine et des patients. Parmi les principes régis par le *serment d'Hippocrate* d'origine, figure celui de la confidentialité des informations, apprises du patient ou de sa famille, ou découvertes à l'occasion de la consultation médicale. [1]

(0a) Ἃ δ᾽ ἂν ἐν θεραπείῃ ἢ ἴδω, ἢ ἀκούσω, ἢ καὶ ἄνευ θεραπηίης κατὰ βίον ἀνθρώπων, ἃ μὴ χρή ποτε ἐκλαλέεσθαι ἔξω, σιγήσομαι, ἄῤῥητα ἡγεύμενος εἶναι τὰ τοιαῦτα.

(0b) Και όσα τυχόν βλέπω ή ακούω κατά τη διάρκεια της θεραπείας ή και πέρα από τις επαγγελματικές μου ασχολίες στην καθημερινή μου ζωή, αυτά που δεν πρέπει να μαθευτούν παραέξω δεν θα τα κοινοποιώ, θεωρώντας τα θέματα αυτά μυστικά.

S'il existe de nombreuses traductions de ce serment, nous en avons retenu deux, relatives à la confidentialité des informations. La première traduction fut donnée par le philosophe et lexicographe français Émile Littré, la seconde est celle renseignée dans le catalogue CISMeF [2] (Catalogue et index des sites médicaux de langue française) [Darmoni et al., 2000] du CHU de Rouen. [3]

(1) « *Quoi que je voie ou entende dans la société pendant, ou même hors de l'exercice de ma profession, je tairai ce qui n'a jamais besoin d'être divulgué, regardant la discrétion comme un devoir en pareil cas.* »

(2) « *Tout ce que je verrai ou entendrai autour de moi, dans l'exercice de mon art ou hors de mon ministère, et qui ne devra pas être divulgué, je le tairai et le considérerai comme un secret.* »

Si les jeunes diplômés de pharmacie et de médecine continuent, à l'heure actuelle, de prononcer le *serment d'Hippocrate*, il s'agit avant tout de sanctionner l'obtention du diplôme et de symboliser la fin des études. Les études en médecine restent cependant régies par le *Code de déontologie médicale*, établi par l'Ordre national des médecins.

1. La version (0a) est en grec ancien, la version (0b) en grec moderne. Source : Wikipedia.
2. http://www.chu-rouen.fr/cismef/
3. http://www.chu-rouen.fr/ssf/art/sermenthippocrate.html

Il n'en demeure pas moins que ces principes de confidentialité et de respect de la vie privée restent d'actualité, à plus forte raison depuis que les données existent sous forme numérique et que les moyens de communication en facilitent les échanges.

Motivations et objectifs

Ainsi émerge la problématique de l'anonymisation des documents cliniques, dès lors que ces documents sont utilisés en dehors du parcours de soins. Que les documents soient rassemblés en corpus à des fins de recherche *(recherche de cas similaires, développement d'outils informatiques, etc.)*, ou utilisés dans des publications scientifiques *(parutions de cas d'étude)*, les informations personnelles contenues dans ces documents ne doivent pas être lisibles par tout un chacun. Un processus d'anonymisation est alors engagé, au terme duquel la réidentification du patient auquel il est fait mention dans le document ne doit plus être possible.

Dans le cadre du présent travail, nous cherchons à étudier les différentes méthodes existantes en traitement automatique des langues (TAL) pour anonymiser, de la manière la plus automatique possible, les documents cliniques disponibles sous forme écrite. Ces documents sont rédigés en langue naturelle *(par opposition aux langages, tels que le langage des animaux ou les langages de programmation en informatique)*, avec des particularités propres au domaine biomédical. Ces particularités font de la langue utilisée dans ces documents une langue de spécialité, avec ses codes et ses usages partagés par la communauté biomédicale.

L'une des principales difficultés qui apparait lorsque l'on travaille en TAL concerne le fait qu'il est difficile de formaliser les langues naturelles (c.-à-d. qu'il est difficile de modéliser le fonctionnement des langues sur ordinateur, contrairement à un programme informatique composé d'une suite d'instructions). Bien que les langues soient régies par des règles *(des règles de grammaire, des règles de prononciation, etc.)*, à chaque règle est associé un certain nombre d'exceptions. À titre d'exemple, la règle de formation du pluriel des substantifs en français par ajout d'un « s » final est enfreinte par plusieurs types d'exceptions : (i) les mots invariables *(une souris/des souris)*, (ii) la transformation d'une désinence au singulier en une version différente au pluriel *(un cheval/des chevaux)*, ou bien encore (iii) l'utilisation de radicaux distincts au singulier et au pluriel *(un œil/des yeux)*. Ainsi, l'objectif global poursuivi par le traitement automatique des langues s'accompagne généralement d'une tentative de formalisation maximale des langues naturelles.

Démarche suivie

Dans ce travail de thèse, nous nous intéressons à la problématique de l'anonymisation automatique de documents en domaine médical. J'ai abordé pour la première fois cette problématique lors de mon stage de DESS *Ingénierie Multilingue* [4] effectué au CHU de la Pitié-Salpêtrière [5] sous la direction de Pierre Zweigenbaum. Ce stage portait sur la constitution d'une chaîne d'anonymisation automatique pour des comptes rendus opératoires et des lettres de suivi en stomatologie/chirurgie maxillo-faciale. Quelques années plus tard, ce travail de thèse s'inscrit dans la continuité de celui entrepris à la Pitié-Salpêtrière. Il repose sur l'utilisation et la combinaison de plusieurs méthodes permettant de traiter les documents écrits du domaine médical rédigés en langue naturelle et disponibles sous forme électronique.

L'objectif poursuivi dans ce travail concerne la recherche des techniques permettant l'anonymisation automatique de documents cliniques avec un double impératif : celui de la meilleure couverture possible des anonymisations tout en préservant au maximum la qualité d'origine du document de manière à permettre l'utilisation du corpus ainsi anonymisé dans un processus de recherche ultérieur. Sachant qu'il existe plusieurs méthodes informatiques pour traiter automatiquement les documents rédigés en langue naturelle, ce travail vise à identifier quelles sont les techniques les plus efficaces pour répondre à notre problématique. Ce travail de thèse s'accompagne de la production de plusieurs éléments : un guide d'annotation, un corpus de référence manuellement annoté et un outil d'aide à l'anonymisation automatique.

En premier lieu, nous avons établi un guide d'annotation dont l'objectif est de constituer un corpus annoté de référence. Ce guide fixe les principes d'annotation, présente des exemples et définit les douze catégories d'information à anonymiser que nous avons retenues. Dès la création de ce guide se pose un certain nombre de questions, principalement sur les catégories d'informations qu'il importe de traiter d'une part, et sur les frontières de l'anonymisation d'autre part *(quels éléments doivent faire partie de la portion annotée ?)*.

Sur la base de ce guide d'annotation, nous avons manuellement annoté un corpus de 562 comptes rendus cliniques en cardiologie. Parmi ces documents, 100 ont fait l'objet d'une double annotation, avec calcul d'accord inter-annotateurs *(coefficient κ)*, suivie d'une phase d'adjudication des annotations. Ce corpus de référence a été scindé en trois : un corpus d'entraînement servant de base à l'apprentissage, un corpus de développement pour optimiser la construction du modèle statistique et un corpus de test pour évaluer les résultats des systèmes.

Disposant d'un corpus de référence d'une qualité jugée satisfaisante, nous avons alors orienté notre travail vers l'étude des méthodes existantes et l'implémentation d'outils relevant de ces différentes méthodes pour effectuer le travail d'anonymisation automatique. Nous nous sommes ainsi intéressé aux deux grandes familles

4. INaLCO, http://www.inalco.fr/, Institut National des Langues et Civilisations Orientales *(« Langues'O »)*, Paris.

5. AP–HP (Assistance Publique – Hôpitaux de Paris), DSI (Direction des Systèmes d'Information), STIM (mission de recherche en Sciences et Technologies de l'Information Médicale).

de méthodes disponibles en TAL, les méthodes symboliques – à base de règles et de listes – et les méthodes par apprentissage statistique. Nous avons ainsi testé chacune de ces méthodes isolément, puis en combinaison l'une avec l'autre. À cet effet, nous avons produit un outil d'aide à l'anonymisation automatique des données personnelles pour chacune des deux méthodes testées.

La première contribution de ce travail de thèse consiste en l'étude et l'expérimentation des deux grandes méthodes existantes pour anonymiser des informations personnelles dans le domaine clinique, en appliquant les outils que nous avons développés sur le même jeu de données. Une deuxième contribution concerne l'impact des définitions humaines retenues en matière de catégorisation des informations traitées *(distinction nom/prénom, définition des adresses sans distinguer les composants)* au regard des performances des anonymisations réalisées globalement et pour chacune de ces catégories. Enfin, la dernière contribution consiste à situer dans le processus global de l'anonymisation le traitement appliqué aux informations les plus sensibles *(noms et prénoms)* selon la méthode choisie.

Plan de lecture

Ce mémoire de thèse est divisé en trois parties, précédées d'un chapitre posant la problématique abordée dans ce travail : le contexte général dans lequel s'inscrit ce travail en termes d'état de l'art et de modalités d'évaluation (partie I) puis les expériences réalisées au moyen de plusieurs méthodes (partie II). Nous donnons dans les annexes (partie III) un aperçu du travail réalisé au travers des manuels techniques produits.

Problématique. Dans le chapitre 1, nous présentons le cadre théorique général dans lequel s'inscrit ce travail de thèse. À cet effet, nous introduisons le sujet avec une présentation générale des corpus de manière à introduire les besoins existants en matière d'anonymisation automatique de comptes rendus cliniques (section 1.2). Nous brossons ensuite un portrait de la terminologie employée autour de l'anonymisation en langue générale et appliquée aux corpus médicaux (section 1.3), avant de présenter une catégorisation des informations à anonymiser selon les types de corpus (section 1.4). Nous poursuivons sur le processus d'anonymisation indispensable aux projets de recherche scientifique travaillant sur des données cliniques (section 1.5). Enfin, nous abordons une réflexion sur la possibilité de disposer d'un outil d'anonymisation qui soit multilingue, multi-culturel et multi-disciplinaire (section 1.6).

PARTIE I. ÉTAT DE L'ART. La première partie de ce mémoire introduit le sujet en deux chapitres.

Méthodologies. Dans le chapitre 2, nous présentons les différentes méthodologies existantes pour effectuer une anonymisation de manière automatique. Nous regroupons les différentes méthodes selon le processus utilisé. L'utilisation de méthodes symboliques (section 2.2), l'utilisation de méthodes par apprentissage statistique (section 2.3) avec une présentation rapide des principaux formalismes utilisés, et finalement, l'hybridation des deux derniers types de méthodes (section 2.4).

Évaluation. Dans le chapitre 3, nous détaillons l'évaluation dans le domaine du traitement automatique des langues appliquée à la problématique de l'anonymisation automatique. Nous introduisons les différentes mesures d'évaluation employées dans le TAL (section 3.2) en discutant pour chacune d'entre elles de la possibilité de les utiliser pour évaluer un processus d'anonymisation. Nous présentons par ailleurs les possibilités de biais ou de relâchement de contraintes permettant de traiter de manière optimale l'anonymisation des données personnelles. À l'opposé de ces évaluations automatiques existent les évaluations humaines, parfois nécessaires, en particulier pour vérifier qu'aucune information sensible ne demeure dans un corpus avant sa redistribution hors du parcours de soin (section 3.3). Nous introduisons également les coefficients d'accord inter-annotateurs utilisés pour évaluer la qualité des annotations humaines, nécessaires pour la constitution des corpus de référence qui seront utilisés pour effectuer l'évaluation automatique (section 3.4). Nous terminons par les mesures qui nous permettent d'accorder du sens aux résultats produits par un système, en termes de d'intervalles de confiance (section 3.5).

PARTIE II. EXPÉRIMENTATIONS. Dans la deuxième partie, nous détaillons les expériences que nous avons menées en matière d'anonymisation automatique.

Corpus et matériau. Dans le chapitre 4, nous présentons les travaux préparatoires aux différentes expériences d'anonymisation. Ces travaux concernent la rédaction d'un guide d'annotation fixant les principes que doivent suivre des annotateurs humains pour produire un corpus de référence (section 4.2). Nous détaillons par la suite les caractéristiques du corpus de cardiologie sur lequel nous avons travaillé, les objectifs du projet dans lequel s'intègre ce travail, et le processus de double annotation par des humains pour produire le corpus de référence (section 4.3). Nous terminons ce chapitre par une présentation des outils utilisés pour annoter le corpus, évaluer les résultats et interpréter les résultats produits (section 4.4).

Méthodes symboliques. Dans le chapitre 5, nous détaillons les expériences que nous avons menées à base d'approches symboliques. Dans un premier temps, nous introduisons nos premières expériences d'anonymisation réalisées en 2002 sur un corpus en stomatologie et dont nous nous sommes inspiré (section 5.2). Dans un second temps, nous présentons deux expériences que nous avons menées pour anonymiser un corpus en cardiologie. La première concerne la tentative d'adaptation au français d'un outil existant et l'explication, par une analyse des erreurs produites, des raisons qui nous ont conduit à ne pas poursuivre cette adaptation jusqu'à son terme (section 5.3). La deuxième expérience se rapporte au développement d'un nouvel outil, pour lequel nous présentons les caractéristiques techniques et l'évaluation réalisée (section 5.4).

Méthodes par apprentissage. Dans le chapitre 6, nous présentons nos expériences portant sur les approches par apprentissage statistique. Dans un premier temps, nous détaillons le protocole expérimental que nous avons défini et suivi dans ces expériences (section 6.2). Après avoir exposé les différents outils à notre disposition, nous justifions celui que nous avons retenu et les paramètres de configuration que nous avons choisis (section 6.3). Enfin, nous présentons les expérimentations

que nous avons menées (section 6.4) en exposant, d'une part les caractéristiques produites pour nourrir la construction du modèle, et d'autre part la mise en place des différentes expériences que nous avons réalisées.

Évaluation et discussion. Dans le chapitre 7, nous rapportons les résultats obtenus sur les différentes expériences menées, d'abord sur les méthodes symboliques (section 7.2), puis sur les approches à base d'apprentissage statistique (section 7.3), et enfin sur l'enchaînement des deux méthodes (section 7.4). Nous présentons par la suite une analyse des erreurs produites lors de la meilleure expérience (section 7.5) et discutons des résultats obtenus, notamment du point de vue de la complémentarité et des différences entre les approches suivies (section 7.6). Enfin, nous dressons un bilan du traitement des informations dites « sensibles », en l'occurrence les noms et prénoms, au terme des différentes expériences d'anonymisation réalisées (section 7.7).

PARTIE III. ANNEXES TECHNIQUES. La dernière partie regroupe les annexes dans lesquelles nous avons rassemblé certains des éléments produits dans ce travail.

Guide d'annotation. Nous reproduisons dans l'annexe A le guide d'annotation que nous avons élaboré pour annoter manuellement un corpus de comptes rendus hospitaliers en cardiologie. Dans ce guide, nous présentons les objectifs poursuivis, nous listons les éléments à anonymiser et les principes d'anonymisation qui doivent être suivis. Nous détaillons enfin les catégories définies au moyen d'exemples issus du corpus. Ce guide a été utilisé par deux annotateurs humains pour produire un corpus de référence sur lequel ont été évalués les systèmes produits.

Medina. Dans l'annexe B, nous fournissons le manuel d'utilisation du système d'anonymisation par méthodes symboliques intitulé « Medina ». Ce manuel présente rapidement les modalités de lancement de l'outil et les options disponibles selon le type de sortie attendue. Nous fournissons un exemple de l'anonymisation produite au terme de chaque étape, ainsi qu'un historique de la constitution de cet outil.

Medina-CRF. L'annexe C est constituée du manuel d'utilisation de la version par apprentissage intitulée « Medina-CRF ». Ce manuel reprend les principes d'utilisation de la chaîne de traitements, un exemple d'utilisation pas à pas et l'historique.

Configuration. Dans l'annexe D, nous reproduisons le fichier de configuration utilisé par l'outil CRF qui nous a donné les meilleurs résultats dans nos expériences.

Chapitre 1

Problématique

Le soleil se leva lentement, comme
s'il doutait de l'utilité de cet effort.

Le huitième sortilège
TERRY PRATCHETT

Sommaire

1.1 Introduction

Dans ce chapitre, nous abordons la problématique traitée dans ce travail de thèse en présentant le cadre théorique dans lequel nous nous inscrivons.

Nous abordons cette présentation du point de vue de l'anonymisation de documents cliniques en présentant les corpus de textes médicaux, du point de vue de leurs caractéristiques et de leur utilité. Ayant identifié l'intérêt des corpus médicaux, nous introduisons la problématique de l'anonymisation automatique des documents cliniques, en particulier dans le processus de recherche scientifique fondée sur la découverte de connaissances en corpus.

Dans un premier temps, nous définissons les termes qui gravitent autour du concept d'« anonymisation » en langue générale, puis appliqués aux corpus médicaux, aussi bien en français qu'en anglais.

Nous présentons ensuite une typologie des catégories d'informations devant faire l'objet d'une anonymisation. Trois catégories principales se dégagent : les informations nominatives et numériques, les informations préjudiciables, et les informations combinées.

Enfin, nous concluons ce chapitre par une réflexion sur la possibilité de disposer d'un outil d'anonymisation qui soit multilingue, multi-culturel et multi-disciplinaire.

1.2 Les corpus de textes médicaux

1.2.1 Typologie des corpus

Au niveau médical, on distingue deux principaux types de documents, aux contenus et aux modalités de constitution assez opposés en raison de la finalité de ces documents : (i) les articles scientifiques, et (ii) les documents cliniques qui composent le dossier médical. Ces deux types recouvrent des réalités bien différentes. La constitution de corpus diffère largement d'un type à l'autre, tant en termes d'accès — il est plus facile de récupérer des résumés dans MEDLINE que d'accéder à des comptes rendus cliniques — que de traitements nécessaires et de contraintes juridiques qui pèsent sur leur diffusion, notamment en matière d'anonymisation. En ce qui concerne ce travail de recherche, nous ne nous intéressons qu'au deuxième type de corpus, le dossier médical. En effet, tout élément clinique à paraître dans un article scientifique aura déjà fait l'objet d'une anonymisation.

Les corpus d'articles scientifiques

Le premier type de documents se compose d'articles scientifiques rédigés par les médecins et chercheurs du domaine biomédical, ou de résumés d'articles scientifiques indexés dans la base de données bibliographique MEDLINE. [1] Il s'agit donc de docu-

1. MEDLINE est une base de données qui rassemble les références bibliographiques et les résumés de la littérature biomédicale du monde entier. Les articles présents dans la base sont indexés avec des termes extraits du thésaurus biomédical MeSH *(Medical Subject Headings)*, http://www.nlm.nih.gov/mesh/. Une interface sur internet, nommée PubMed, permet d'accéder à MEDLINE. Elle fournit, lorsque cela s'avère possible, des liens vers les articles complets, http://www.ncbi.nlm.nih.gov/

ments scientifiques, rédigés par et pour des chercheurs, sur des thématiques de recherche précises. Parmi les différents types d'articles figurent les articles d'analyses *(« review »)* et ceux présentant un cas d'étude particuliers *(« case report »)*. Ils sont représentés dans MEDLINE par une liste de type de publications *(« publication type »)*.

Chaque article présent dans la base de données est accompagné de méta-données *(noms, prénoms et initiales des auteurs, affiliation du premier auteur, type de publication, nom complet et abréviation ISO*[2] *de la revue ou de la conférence, numéro ISSN, titre, résumé, langue de publication, année, volume, numéro, pagination, identifiant PubMed, DOI,*[3] *date de publication en ligne, date d'intégration dans* MEDLINE*, indexation MeSH)* renseignées par les personnes indexant les articles d'après les informations disponibles dans les articles, ou, de plus en plus, fournis par les éditeurs scientifiques.

Les corpus de documents cliniques : le dossier médical

Présentation. Le second type de corpus rassemble des documents cliniques, dans le sens où ces documents contiennent des données relatives aux patients. L'ensemble de ces documents constitue le dossier médical du patient dont la rédaction est obligatoire, quel que soit le statut de l'établissement de soins, public ou privé,[4] ou le statut du professionnel de santé, tel le praticien libéral.[5] Ce dossier comprend plusieurs types de documents, rédigés à des moments divers du séjour hospitalier, et pour des destinataires distincts. Il intègre trois types d'informations : (i) les informations recueillies lors de consultations externes, lors de l'accueil aux urgences, lors de l'admission et au cours du séjour hospitalier, et (ii) les informations formalisées établies à la fin du séjour. Un troisième type rassemble les informations obtenues auprès de tierces personnes et qui ne concernent pas la prise en charge thérapeutique.

Le dossier médical intègre obligatoirement[6] des informations nominatives : l'identité du patient ou celle de la personne de confiance *(nom, prénom, date de naissance ou numéro d'identification)*, et l'identification de la personne à prévenir. Le dossier mentionne également l'identité du professionnel de santé qui a recueilli ou produit les données. D'autre part, les prescriptions médicales doivent être datées avec indication de l'heure, et signées. Comme n'importe quel document technique, la rédaction d'un document médical s'apprend et obéit à des contraintes de présentation et d'information *(renseignements administratifs, médicaux, structure syntaxique des phrases employées, etc.)*. Le rédacteur du compte rendu hospitalier (CRO) doit ainsi vérifier les informations contenues, sa responsabilité étant engagée.

De plus en plus, les professionnels de la santé sont invités à produire le dossier médical sous forme électronique. Le fait que le dossier existe sous forme électronique suppose préalablement une déclaration auprès de la CNIL. Le dossier médical sous

pubmed. Les outils MEDLINE, MeSH et PubMed sont fournis par la NLM *(National Library of Medicine)* aux États-Unis.

2. Les éléments des noms de revue sont généralement abrégés comme suit : Am=American, Annu=Annual, Assoc=Association, Biomed=Biomedical, Inform=Informatics, Int=International, J=Journal, Med=Medical, Proc=Proceedings, Stud=Studies, Symp=Symposium, Technol=Technology. Ainsi, le *Journal of the American Medical Informatics Association* sera abrégé *J Am Med Inform Assoc*.

3. DOI *(Digital Object Identifier)*, système d'identification pérenne des objets numériques. Un identifiant est attribué de manière unique à chaque objet *(dont les articles scientifiques)* et peut être décodé en passant par l'interface http://dx.doi.org/.

4. Article R1112-2 du Code de la santé publique.

5. Article 45 du Code de déontologie médicale.

6. Article R1112-3 du Code de la santé publique.

forme électronique présente les avantages suivants : formatage pré-défini du document avec des zones de saisie assistée et des zones de texte libre, accès facilité depuis n'importe quel poste de travail, accès plus rapide entre intervenants d'un même établissement *(le dossier est disponible après l'intervention sans délai d'édition ou de transmission)*, outil d'indexation *(avec codage PMSI)* et de recherche, décomptes statistiques, etc. Les propriétés du dossier en termes de données cliniques *(ayant une valeur juridique en cas de procès)* sont préservées. À l'opposé, le support électronique et les facilités d'échange impliquent une attention accrue quant au maintien du secret professionnel : « *Le médecin doit protéger contre toute indiscrétion les documents médicaux concernant les personnes qu'il a soignées ou examinées, quels que soient le contenu et le support de ces documents.* ».[7]

Consultations externes, accueil, admission, séjour hospitalier. Les informations relevant du premier type ont été produites avant et pendant le séjour hospitalier. Elles sont rassemblées dans les types de documents suivants :

- Les correspondances entre professionnels de la santé, notamment la lettre du médecin traitant à l'origine de l'hospitalisation et la lettre initiale pré-opératoire *(certificat médical initial, CMI)* ;
- Le consentement du patient si la situation juridique le requiert ;
- Le dossier d'anesthésie ;
- Le compte rendu opératoire (CRO) ou d'accouchement ;
- La liste des actes transfusionnels pratiqués et l'éventuelle fiche d'incident transfusionnel ;
- Les résultats de laboratoire *(analyse sanguine, anatomopathologie)* ;
- Les fichiers d'imagerie médicale *(échographie, radiographie, scintigraphie)* ;
- Le dossier de soins infirmiers.

Dans ces documents, on trouvera ainsi plusieurs catégories d'informations cliniques : (i) les motifs d'hospitalisation, (ii) la recherche d'antécédents et de facteurs de risques, (iii) les conclusions de l'évaluation clinique initiale, (iv) le type de prise en charge prévu et les prescriptions effectuées à l'entrée, (v) la nature des soins dispensés et les prescriptions établies lors de consultations externes ou lors du passage aux urgences, (vi) les informations liées à la prise en charge hospitalière *(état clinique, soins reçus, examens para-cliniques, imagerie, etc.)*, (vii) toutes les informations de prescription médicale, d'exécution de la prescription et d'examens complémentaires, et $(viii)$ n'importe quelle autre information clinique produite par des professionnels de la santé.

Ces documents intègrent des informations administratives directement identifiantes, ou indirectement par le biais de recoupements *(nom de l'établissement, nom, prénom, date de naissance, adresse du patient, date de l'intervention, noms des professionnels de santé — opérateur, anesthésiste, infirmière, aide opératoire —, cotation CCAM*[8]*)* et des informations cliniques *(diagnostic pré-opératoire, choix thérapeutique, description de l'état pathologique et des lésions, incidents, implants, drainages, prélèvements anatomopathologiques et bactériologiques, pansements et contentions réalisés, durée de l'intervention, consignes post-opératoires)*. La connaissance de ce type d'information dans les documents cliniques est essentielle pour notre

7. Article 73 du Code de déontologie médicale.

8. CCAM *(Classification Commune des Actes Médicaux)*, nomenclature établie pour coder les différents actes médicaux et définir les honoraires de ces actes, http://www.ameli.fr/accueil-de-la-ccam/index.php

travail d'anonymisation.

Fin du séjour hospitalier. Le deuxième type d'information rassemble les informations cliniques produites pendant le séjour hospitalier. Elles sont formalisées à l'occasion de la sortie de l'établissement de soins en différents types de documents :

- Le compte rendu d'hospitalisation (CRH) ;
- La lettre de sortie *(à destination du médecin traitant).* Elle mentionne les incidents rencontrés et les conseils de surveillance ou précautions à observer ;
- Les doubles de l'ordonnance de sortie ;
- La fiche de liaison infirmière.

Ces documents rassemblent ainsi les informations cliniques établies lors du séjour hospitalier. Ils intègrent également des informations relatives au séjour post-hospitalier : (i) les prescriptions de sortie, et (ii) les modalités de sortie *(séjour à domicile ou dans d'autres structures de soin).*

Exemple. Nous donnons ci-dessous (voir tableau 1.1) un exemple de compte rendu hospitalier provenant d'un service de cardiologie d'un CHU français. Les informations identifiantes *(noms, prénoms, dates et lieux)* ont été remplacées par de fausses informations pour les besoins de la démonstration. Dans ce CRH, on retrouve ainsi les principales catégories d'informations attendues : l'histoire de la maladie associée au motif de l'hospitalisation, les analyses effectuées, les observations réalisées, le traitement de sortie et les conseils post-opératoires.

<texte>Cher Ami,

Monsieur Michael Stipe (04.01.60) est malheureusement revenu dans le service du 28 avril au 5 mai 1993 pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en juin 92.

Cet infarctus est survenu alors que le patient était resté depuis lors asymptomatique avec des épreuves d'effort de bon niveau et négatives. L'infarctus a été vu en cardiologie à la sixième heure. Le contrôle coronarographique effectué à l'entrée révélait une occlusion de la coronaire droite à la jonction des segments I et II en un site où la coronarographie de juin 92 ne révélait qu'une irrégularité minime. L'artère a été rapidement désobstruée. Le pic de CPK a atteint 1957 le lendemain avec des CPK à 110. L'évolution a été simple. L'échocardiographie effectuée dès le 29 avril (P. BUCK) retrouvait une hypokinésie septo-apicale avec dyskinésie apicale localisée, et une akynésie postéro-inférieure, une dilatation moyenne du VG avec une altération de la fonction systolique globale : la fraction d'éjection ventriculaire gauche étant évaluée à 40 %, ce qui était retrouvé lors de la sortie d'hospitalisation en juillet. Il existe un possible petit thrombus apical mural, et par ailleurs une fuite mitrale modérée. L'épreuve d'effort effectuée au sixième jour a permis de soutenir la charge de 150 W pendant 3 mn et d'atteindre la fréquence cardiaque de 133/mn (sous bêta-bloquant). Elle est restée négative cliniquement et électriquement. Il s'agit d'un très bon niveau d'effort superposable aux résultats des tests précédents. L'enregistrement Holter des 24 H ne retrouve pas d'arythmie significative. L'ECG haute amplification ne retrouve pas de potentiel tardif ventriculaire.

Il semble donc que ce deuxième infarctus n'ait entrainé que des dégâts myocardiques limités. Cependant, la survenue de ce deuxième infarctus reste très surprenante. Il survient moins d'un an après un infarctus dans un autre territoire, alors que l'artère coronaire droite en cause semblait très peu pathologique l'an dernier. Le bilan des facteurs de risque chez ce patient est négatif (les résultats concernant les dosages d'homocystéine et des vitamines B impliquées dans son métabolisme sont revenus également négatifs). Nous avons effectué une recherche assez complète des facteurs d'hémostase dont nous attendons les résultats. Toutefois, on sait que la coronarographie peut sous estimer l'existence de lésions athéromateuses réelles et il est vraisemblable que la coronarographie de juin 92 sous estimait une probable lésion (vulnérable) au niveau de la coronaire droite, vis à vis de laquelle les mesures préventives instituées au décours du premier infarctus (conseils alimentaires et réadaptation cardiovasculaire) n'ont pas eu le temps d'être actives. Nous avons donc tenté de rassurer Monsieur Stipe à cet égard en espérant que ces mesures puissent donner pleinement leur effet dans les années à venir. Le bilan lipidique recontrôlé il est vrai au décours de l'infarctus reste strictement normal (cholestérol total 3,8 mmol/l ; triglycérides 1,09 mmol/l ; HDL-c 0,96 mmol/l ; LDL-c 2,35 mmol/l). Nous avons seulement adjoint à son traitement un peu de vitamine E sous forme de TOCO 500 1/jour.

Le traitement de sortie associe :
- TENORMINE 1 cp/jour,
- TRIATEC 5 mg 1/jour,
- SOLUPSAN 160 mg/jour,
- TOCO 500 1/jour,
- TICLID 2/jour, à poursuivre pendant 1 mois en raison de l'angioplastie (avec la surveillance hématologique régulière nécessaire).

Par ailleurs, compte tenu du possible petit thrombus mural, nous envisageons de placer Monsieur Stipe sous anti-vitamines K lorsque la période d'association Aspirine - TICLID motivée par l'angioplastie sera terminée (1 mois environ). Auparavant, nous recontrôlerons très prochainement l'échocardiographie.

Nous restons à ta disposition en cas de difficulté.

Avec nos salutations les plus cordiales.

M. MILLS Docteur B. BERRY
Interne Salle Athens Praticien Hospitalier.
</texte>

TABLE 1.1 – Exemple de compte rendu hospitalier

1.2.2 Utilité des corpus

Les corpus de textes constituent une ressource importante en traitement automatique des langues. Ils se révèlent nécessaires, tout en soulevant de nombreux problèmes techniques.

Développement d'outils. En premier lieu, ils permettent de développer de nouveaux systèmes informatiques ou d'adapter des systèmes existants. Ces créations et adaptations de systèmes sont effectuées pour permettre de couvrir de nouvelles réalités présentes dans le corpus ou, plus généralement, pour mieux appréhender les spécificités linguistiques d'une nouvelle thématique ou d'un nouveau domaine.

Entraînement d'outils. En second lieu, lorsqu'ils sont porteurs d'informations complémentaires et d'annotations, les corpus sont utiles pour entraîner des systèmes reposant sur des méthodes par apprentissage. Les systèmes reposant sur un apprentissage construisent des modèles sur la base d'observations faites en corpus, de manière à modéliser une tâche *(classification, indexation, typage, etc.)*. Pour être efficace, ce type de méthode doit bénéficier d'un volume important d'annotations en corpus, de manière à couvrir un maximum de situations *(contextes et indices « internes »)* dans lesquels les annotations ont été réalisées et que le modèle va chercher à modéliser.

Évaluation de performances. Enfin, les corpus sont indispensables pour évaluer des systèmes sur des tâches précises. Les campagnes d'évaluation reposent ainsi sur le principe d'une confrontation de plusieurs systèmes sur un même jeu de données *(un corpus annoté servant de référence)* de manière à permettre une évaluation des différentes méthodes et approches suivies par chaque participant. D'une édition à l'autre, les campagnes d'évaluation permettent de suivre la progression d'un même système dans le temps. Les campagnes d'évaluation présentent deux avantages majeurs : (i) elles donnent lieu à des avancées notables en matière de recherche et développement *(création ou amélioration d'outils existants, découverte de nouvelles pistes d'exploitation de ressources existantes, etc.)* ; et (ii) elles offrent l'opportunité de créer des ressources *(corpus annoté)* qui peuvent être redistribuées gratuitement auprès de la communauté scientifique.

Intérêt des corpus cliniques. L'utilisation qui peut être faite de ces corpus de documents cliniques est particulièrement variée. Dans le cadre des campagnes d'évaluation, de tels types de corpus ont été utilisés pour de nombreuses tâches :

- La détection de gènes et de leurs relations dans la littérature scientifique (tâche réalisée à partir de 2 000 résumés indexés dans MEDLINE) [Kim et al., 2003] ;
- L'indexation et le codage de documents d'après les codes de la terminologie médicale CIM-9 [9] (évaluation BioNLP 2007) [Pestian et al., 2007] ;
- L'identification des prescriptions médicamenteuses (évaluation i2b2 2009, noms de médicaments et informations associées : *dosage, mode d'administration, fréquence, durée, raison de la prescription*) [Uzuner et al., 2010] ;

9. La CIM-9 (Classification Internationale des Maladies, 9e révision), *International Classification of Diseases, 9th Revision, Clinical Modification* (ICD-9-CM), est une classification des maladies et traumatismes. Elle est utilisée en milieu hospitalier pour traiter de la morbidité hospitalière. La terminologie comprend dix-sept chapitres segmentés en sections, catégories, et sous-catégories sous lesquels sont classées les maladies. Un code unique à trois chiffres est associé à chaque maladie. Le manuel de codage est accessible à l'adresse suivante : http://www.health.fgov.be/internet2Prd/groups/public/@public/@dg1/@datamanagement/documents/ie2divers/994482_fr.pdf. La CIM-9 est la version actuellement utilisée aux États-Unis. Elle a été remplacée par la CIM-10 *(composée de 21 chapitres)* dans les États membres de l'OMS en 1994, les États-Unis continuant d'utiliser la CIM-9. L'adoption de la CIM-10 aux États-Unis, initialement prévue pour le 1er octobre 2013, a été repoussée d'une année.

- L'identification et le typage de concepts médicaux *(examen, problème, traitement)*, les assertions sur ces concepts *(absent, conditionnel, hypothétique, possible, présent, associé à quelqu'un d'autre)* et les relations entre concepts *(examen qui diagnostique/révèle un problème, problème qui indique un autre problème, traitement qui améliore/dégrade/cause un problème, traitement administrable/non administrable pour un problème)* (évaluation i2b2/VA 2010) [Uzuner et al., 2011] ;
- La résolution de coréférences entre concepts médicaux (*examen, personne, problème, pronom, traitement*, évaluation i2b2/VA 2011) [Uzuner et al., 2012] ;
- Ou plus récemment, l'identification des relations temporelles entre concepts médicaux *(département clinique, examen, occurrence, preuve, problème, traitement)* et expressions temporelles Timex3 (évaluation i2b2/VA 2012).

De manière plus générale, ces corpus — dans leur globalité — présentent un intérêt particulier en recherche clinique *(découverte de cas similaires)* et translationnelle *(amélioration des soins apportés aux patients par la découverte d'information innovante)*, ainsi qu'en ingénierie *(développement et maintenance d'outils)*. Les informations contenues dans ces documents peuvent servir de support à des articles scientifiques *(publication d'études de cas)*. Sur le plan linguistique, ces corpus mettent en évidence les caractéristiques *(choix terminologiques, constructions morphologiques, tournures syntaxiques, etc.)* d'une langue de spécialité, la langue médicale, dans des types de discours particuliers *(compte rendu médical, lettre au médecin traitant, etc.)*. L'étude linguistique de ces particularités nécessitent de disposer de corpus anonymisés représentatifs de cette langue médicale [Grabar, 2004].

1.2.3 La langue médicale, une langue de spécialité

Présentation. Tous les documents du dossier médical sont rédigés par des professionnels de santé et sont destinés à d'autres professionnels de la santé, voire au patient lui-même. Ils obéissent à des normes de rédaction. À l'instar de la théorie de Sapir-Whorf qui établit que la langue conditionne en partie la manière dont on perçoit le monde [Sapir, 1921, Sapir, 1929, Whorf, 1940, Whorf, 1956], sur le plan linguistique, il a été démontré [Harris, 1968, Kittredge et Lehrberger, 1982] que la langue médicale peut être envisagée comme une langue de spécialité disposant de ses propres codes et caractéristiques. La langue médicale dispose également de son propre vocabulaire. Si des outils génériques de traitement de la langue peuvent s'appliquer au traitement de la langue médicale *(étiqueteur, lemmatiseur, racinisateur)*, ce traitement nécessite cependant des ressources dédiées *(lexiques spécifiques, adaptation des règles de REN)* [Bossy et al., 2012].

Sur l'anglais, [Pestian et al., 2007] mentionnent un certain nombre de ces caractéristiques telles que les phrases sans verbe, une sémantique particulière propre au domaine pour la ponctuation, ou encore l'utilisation inhabituelle de métonymies. Sur les types de comptes rendus rencontrés, il s'agit de documents rédigés avec une organisation structurée ou semi-structurée des informations dans le document *(un compte rendu hospitalier intègre des sections particulières telles que : histoire de la maladie, traitement de sortie, etc.)*.

Structure stylistique et organisationnelle. Sur le français, et pour un type précis de compte rendu — le compte rendu opératoire —, des guides d'aide à la rédaction [10] précisent les informations à renseigner et la manière dont ces informations doivent être présentées. Ces guides témoignent d'une volonté de précision et de concision dans la rédaction des CRO [Farfor, 1976].

Sachant que le document doit pouvoir être lu par l'équipe chirurgicale, mais également par le patient ou tout médecin non chirurgien, les principes stylistiques suivants sont préconisés :

- Les phrases doivent être courtes avec une idée par phrase ;
- Placer l'idée forte en début de phrase : « *l'appendice est...* » plutôt que « *en explorant la fosse iliaque droite, on constate que l'appendice est...* » ;
- Le pronom personnel « on » *(pronom de modestie)* est préféré au pronom « nous » réservé aux articles scientifiques ;
- Les verbes sont conjugués au présent, sauf pour les faits antérieurs ou postérieurs à l'opération ;
- Les verbes et les adverbes d'émotion et de jugement sont à éviter : *on déplore, on regrette, on s'étonne, on se félicite, ..., malheureusement, soigneusement, rigoureusement, minutieusement, etc.* ;
- Les procédures chirurgicales nommées par un nom propre sont entièrement rédigées : « *une incision de Pfannestiel* » et non « *un Pfannestiel* » ;
- Seules les abréviations d'unités de mesure sont utilisées et seules les unités de mesure officielles sont autorisées ;
- Préférer la répétition du même terme à l'usage de synonymes ;
- Les énumérations sont autorisées pour réduire la lourdeur des répétitions.

Au niveau organisationnel, le CRO est un document qui décrit l'opération effectuée de manière objective et synthétique. Plusieurs types d'information ne seront mentionnés que s'ils sont originaux par rapport aux pratiques habituelles *(type d'anesthésie, position de l'opéré, voie d'abord, etc.)*. Le CRO comprend deux paragraphes principaux *(les constatations opératoires avec justification des décisions, les gestes effectués et la fermeture)*. Seront précisées en fin de document les informations absentes du reste du document *(appareillage externe, prélèvements anatomopathologiques et bactériologiques)*.

1.3 L'anonymisation

1.3.1 Définition en langue générale

Notions en français

En français, *Le nouveau Petit Robert* [Rey-Debove et Rey, 1993] propose trois termes pour la notion d'*anonymisation* : le nom « anonymat », l'adjectif « anonyme » et l'adverbe « anonymement ». On observe ainsi qu'il n'existe, ni le verbe « *anonymiser », ni le nom « *anonymisation ».

Dans l'usage, un locuteur du français interprètera le terme « anonymisation »

10. http://umvf.univ-nantes.fr/chirurgie-generale/enseignement/compterendu/site/html/cours.pdf, présentation de la méthodologie de rédaction d'un compte rendu opératoire : principes généraux, style et plan.

comme étant *le processus de rendre quelque chose anonyme*. Il apparaît de prime abord difficile d'utiliser ce substantif avec un objet humain :

- *Je travaille sur l'anonymisation de ce document.*
- **Je travaille sur l'anonymisation de cette personne.*

À défaut de définir le terme « anonymisation », les dictionnaires et thésaurus actuels définissent les termes voisins « anonyme » et « anonymat » comme suit.

Anonyme. Pour *Le nouveau Petit Robert*, quatre sens particulièrement proches sont renseignés pour cet adjectif :

- « *Dont on ignore le nom, ou qui ne fait pas connaître son nom.* » On trouve l'utilisation de ce terme dans des expressions relatives à des personnes qui accomplissent une action de manière inconnue du public. Ce sens renvoie à une démarche volontaire de cacher son identité, mais avec une perception positive ou neutre : *un don anonyme, une foule anonyme* ;
- « *Dont le, la responsable n'a pas laissé son nom ou l'a caché.* » Ce sens se retrouve dans le même type d'expression que le précédent. Il s'oppose cependant au premier sens par la perception négative qu'il induit : *recevoir une lettre anonyme/un appel anonyme* ;
- Au sens figuré, « *Impersonnel, neutre, sans originalité.* » Ce sens s'emploie pour désigner le style d'une personne ;
- Dans le domaine financier, « *Dont le nom du propriétaire n'est pas connu.* » Ce sens est surtout utilisé dans des expressions semi-figées telles que *Bon du trésor anonyme* (1807) et *Société anonyme*, une « *société par actions qui n'est pas désignée par le nom d'aucun des associés.* »

Dans le thésaurus de [Péchoin, 1999], l'adjectif apparaît logiquement sous les deux mêmes entrées que celles sous lesquelles apparaît le terme « anonymat » ; il est alors associé aux adjectifs « ignoré » et « inconnu ». Il s'applique donc à une personne ou à un objet dont on ignore l'identité.

Anonymat. Le dictionnaire *Le nouveau Petit Robert* fournit la définition suivante : « *État de la personne ou de la chose qui est anonyme.* » Il convient alors de se reporter à la définition de l'adjectif « anonyme » pour mieux saisir le sens de cette définition. En ce qui concerne le *Trésor de la Langue Française informatisé*,[11] l'« anonymat » renvoie à l'« *état d'une personne, d'une chose dont on ignore le nom, l'identité* ».

Le thésaurus français [Péchoin, 1999] propose le terme « anonymat » sous deux entrées : les notions de « Secret » et de « Nom ». Sous la première notion, il est associé au terme « clandestinité ». Il s'agit d'un sens négatif, utilisé pour désigner quelque chose que l'on dissimule. Sous la seconde notion, le terme apparaît plus neutre et renvoie au fait de rendre anonyme un nom.

Notions en anglais

En anglais, le terme « anonymization » semble également absent des thésaurus. Comme pour le français, on y trouve les termes « anonymity » et « anonymous » sous l'entrée « *Nomenclature — Naming* », également avec des connotations neutres, voires négatives : « *pseudonymous, soit-disant, self-styled; nameless, anonymous; né[e]* » [Morehead, 2001].

11. http://atilf.atilf.fr/, Trésor de la Langue Française informatisé (TLFi).

Conclusion

Pour conclure sur ces définitions en langue générale, on retiendra que l'anonymat concerne l'état d'une personne ou d'un objet qui est anonyme, et qu'est anonyme ce dont on ignore l'identité. L'anonymisation consistant à rendre anonyme un objet.

1.3.2 Définition appliquée aux corpus

L'anonymisation en corpus concerne généralement le fait de masquer les informations identifiantes d'un individu tout en laissant en clair les autres informations présentes dans le document. C'est autour de cette problématique que nous inscrivons le cadre de ce travail et les réalisations que nous avons produites.

Une deuxième acception repose sur l'anonymisation complète de l'ensemble du corpus au moyen d'un chiffrement sécurisé. L'objectif visé par le chiffrement des données consiste à rendre le corpus totalement illisible par quiconque ne disposant pas de la clé de chiffrement. [Landi et Rao, 2003] ont ainsi défini une méthode d'encryptage des données au moyen de clés privées et publiques fondées sur RSA, pour répondre aux menaces d'attaques bioterroristes et de piratages informatiques.

Notions en français

Bien que le terme français admis soit « anonymisation », au plan légal, la problématique qui doit être résolue est celle de « *l'impossibilité d'identifier des personnes* » [Baude, 2006, p. 77]. L'objectif visé par une procédure d'anonymisation de documents, quel qu'en soit le support *(documents écrits ou oraux)* et le domaine thématique *(documents cliniques, judiciaires, ou tout venant)*, repose sur l'impossibilité d'effectuer une réidentification, autrement dit, l'impossibilité de retrouver l'identité d'une personne sur la base des éléments qui auront été laissés en clair dans le document [Meystre et al., 2010]. Doivent ainsi être anonymisées toutes les informations qui, utilisées seules ou en complément d'autres informations, permettent d'identifier le patient mentionné dans le document.

Dans leur présentation des modalités de constitution d'un corpus d'interactions humaines, [Reffay et Teutsch, 2007] décrivent la tâche d'anonymisation au moyen de la terminologie suivante :

- L'anonymisation : la procédure qui consiste à rendre anonymes les documents traités, soit par effacement de l'information *(une suite d'espaces dans un corpus écrit, un « bip » dans un corpus oral)*, soit par remplacement ;
- L'anonymiseur : l'outil qui permet d'effectuer une anonymisation ;
- L'anonymisateur : la personne qui utilise cet outil.

Dans le cadre du présent travail, nous nous proposons de suivre cette terminologie.

Notions en anglais

Alors que la langue française propose le terme « anonymisation » de manière exclusive, l'anglais oppose deux termes aux définitions différentes : *anonymization* (« anonymisation ») et *de-identification* (« désidentification »). [Meystre et al., 2010] précisent ainsi qu'en anglais, le terme « désidentification » renvoie au résultat d'une reconnaissance d'entités nommées. Elle revient à supprimer les éléments appartenant à des catégories prédéfinies *(des identifiants)* telles que les noms, les prénoms,

les lieux, etc. L'objectif visé par la désidentification consiste donc à faire en sorte que les données ne puissent plus être reliées au patient.

À un niveau plus précis, l'« anonymisation » va au-delà de la simple désidentification puisqu'elle est envisagée comme visant à cacher ou supprimer tout élément identifiant. Cette procédure prend ainsi en compte, non seulement les éléments catégorisables *(informations nominatives, numériques, etc.)*, mais également l'ensemble des éléments qui, pris isolément n'auront aucun pouvoir de réidentification, mais qui, utilisés en combinaison avec d'autres éléments du même document ou d'autres documents relatifs au patient, offriront cette capacité de réidentification.

1.3.3 Évaluer les risques de réidentification

Plusieurs études ont traité de l'évaluation des risques de réidentification d'un individu, soit de manière absolue, soit à partir de corpus anonymisés.

Expérimentations aux États-Unis. Sur la base du recensement 1990 des États-Unis, [Sweeney, 2000] a réalisé plusieurs expériences de calcul de réidentification de personnes sur la base de triplets de caractéristiques. L'auteur a ainsi démontré que 87 % de la population américaine peut être identifiée de manière unique si l'on connait le code postal, le sexe, et la date de naissance complète, 53 % peut l'être à partir du lieu d'habitation *(ville, municipalité, ou village)*, du sexe et de la date de naissance complète, et 18 % sur la base du nom du comté, du sexe et de la date de naissance complète. Au niveau médical, l'auteur a pu croiser les données issues de deux corpus *(des données hospitalières publiquement disponibles auprès du IHCCCC*[12] *anonymisées mais comprenant néanmoins des éléments d'information personnelle et la liste des inscrits sur les listes électorales de Cambridge, Massachusetts)* au moyen de trois critères partagés *(code postal, sexe et date de naissance)* pour réidentifier les patients.

Plus récemment, [Golle, 2006] a confirmé cette étude en travaillant sur le recensement américain de l'année 2000. Il a par ailleurs mis en évidence le fait que, pour un même triplet de caractéristiques, le caractère unique de la réidentification est également fonction de la tranche d'âge à laquelle appartient l'individu. Ainsi, sur le triplet code postal, sexe, et date de naissance, 60 % de la population âgée de moins de 50 ans est identifiable de manière unique ; cette proportion monte à plus de 80 % au-delà de l'âge de 75 ans environ.

Aux États-Unis, [Benitez et Malin, 2010] ont réalisé une étude sur l'estimation des risques de réidentification des patients dans le cadre d'attaques informatiques. Cette étude se fonde sur le principe que la réidentification repose sur l'intersection des données laissées en clair dans les documents *(le genre et l'année de naissance)* avec les informations issues d'autres listes *(les listes électorales disponibles pour chaque État)*. Les auteurs évaluent ces risques de réidentification au moyen de trois mesures : (i) le nombre attendu de réidentifications, (ii) la proportion estimée de la population concernée selon l'État, et (iii) le coût financier induit par une procédure de réidentification *(partant du principe qu'une attaque sera lancée si le gain espéré dépasse le coût des dépenses)*. Les résultats calculés par les auteurs diffèrent selon les

12. IHCCCC : Illinois Health Care Cost Containment Council.

États, avec des variations assez importantes. Ainsi, le pourcentage de population vulnérable aux risques de réidentification à partir des listes électorales varie de 0,01 % à Hawaii jusqu'à 36,90 % dans le Missouri. Concernant les coûts de la réidentification, ceux-ci varient d'un coût nul en Caroline du Nord, en Caroline du Sud et dans l'État de New York à 8 267 $ par réidentification dans le New Hampshire et jusqu'à 17 000 $ par réidentification en Virginie Occidentale.

Rappelant que le caractère unique des individus d'une population est couramment utilisé comme mesure d'évaluation du risque de ré-identification d'un patient *(plus un individu est unique, plus le risque de ré-identification le concernant est élevé)*, [Dankar et al., 2012] ont étudié et évalué quatre mesures d'estimation d'unicité *(estimation de Zayatz, estimation binomiale négative, estimateur de Pitman, μ-argus)* qu'ils ont appliquées sur un corpus de données médicales. Il ressort de leur étude qu'aucune mesure d'estimation ne remplit l'ensemble des critères permettant de déterminer l'unicité d'un patient. Les auteurs démontrent cependant que la combinaison de ces différentes mesures dans un arbre de régression permet d'obtenir des résultats convaincants sur le corpus utilisé.

Expérimentations en France. En France, il n'existe à notre connaissance, aucune étude portant sur l'évaluation des risques de réidentification. D'une part, parce que les données médicales ne sont pas diffusées et partagées, et d'autre part, parce qu'il n'existe pas non plus de listes donnant accès à des informations personnelles telles que les listes électorales mentionnées pour les calculs effectués aux États-Unis.

À titre de comparaison, nous simulons un cas de réidentification potentielle pour la France sur la base du croisement de trois critères *(le groupe sanguin, le nom d'une maladie et le sexe)* issus d'un corpus fictif en section 1.4.3. Partant d'une population totale de plus de 64 millions de personnes, le croisement de ces trois critères en sélectionnant des valeurs rares permet de réduire la population d'origine à sept personnes potentielles seulement. Cette simulation ne permet pas de représenter le pourcentage de réidentifications possibles en France mais vise uniquement à présenter quels sont les risques potentiels de réidentification sur la base des informations laissées en clair *(et dans le cas présent, d'informations qui ne peuvent être anonymisées faute de rendre le corpus clinique inexploitable)*.

Dans le domaine de la vie quotidienne, si la mise en place des CV anonymes réduit les écarts d'accès aux entretiens d'embauches, la discrimination se poursuit dans les étapes ultérieures du recrutement [Le Barbanchon, 2012]. Aucune tentative de réidentification n'est effectuée mais le bénéfice de l'anonymisation s'en trouve annulé.

1.3.4 Confidentialité et traitements futurs

Préalablement à l'anonymisation de tout document, plusieurs questions doivent être envisagées. Les réponses apportées à ces questions conditionnent les traitements futurs qui seront appliqués, pour réaliser l'anonymisation, mais également les possibilités d'utilisation du corpus ainsi anonymisé.

Dans un corpus de documents contenant des données privées, quels sont ceux qui doivent faire l'objet d'une anonymisation ? Des résultats de laboratoire font-ils partie des éléments sensibles qu'il convient d'anonymiser, dans quelle mesure des données relatives à une hospitalisation particulière sont-elles sources d'information pour réidentifier un patient ?

À quel moment l'anonymisation doit-elle être effectuée ? Si, en matière de corpus médicaux, les documents d'origine doivent rester exempts de toute mutilation, d'autres formes de documents (corpus audio ou vidéo) peuvent nécessiter une anonymisation lors de la constitution du corpus (un floutage voire la modification de la voix de certains témoins interrogés).

Quels sont les types d'information qui doivent donner lieu à de tels traitements ? Si les données nominatives viennent immédiatement à l'esprit, quel niveau de détail faut-il atteindre pour garantir un anonymat des documents traités ?

Quelles limites se doit-on de fixer au processus d'anonymisation ? Un juste équilibre doit être trouvé entre la recherche de l'absence de réidentification des patients et le maintien des informations nécessaires à l'objet même de la recherche *(recherche de cas similaires par exemple)*. Les numéros de dents mentionnées dans des comptes rendus opératoires de stomatologie constituent-ils des informations sensibles qu'il convient de traiter ?

Il convient de garder à l'esprit ces différentes questions lorsqu'on effectue une anonymisation de corpus ou lorsque l'on développe des architectures logicielles dans le domaine de la santé. Un équilibre doit ainsi être trouvé entre sécurité des informations et maintien de l'interopérabilité des systèmes développés [El Kalam, 2003].

1.3.5 Par quoi remplacer les informations ?

Au terme de l'étape qui aura conduit l'anonymiseur à traiter les documents, les informations identifiées comme devant être anonymisées devront être masquées. Se pose alors la question du traitement à appliquer à ces informations, autrement dit, par quoi doit-on remplacer les informations anonymisées ? Cette étape de masque se doit d'être réalisée en toute fin de chaîne, juste avant la mise à disposition des corpus.

Utilisation de balises

Un premier moyen simple pour masquer les informations personnelles consiste à remplacer les informations, soit (i) par une suite de caractères *(plusieurs caractères "X" ou un blanc typographique)*, soit (ii) par des balises XML avec conservation du type d'information anonymisée *(la balise <prénom> pour tous les prénoms du document, ou des balises <patient> et <médecin> selon le rôle de la personne anonymisée)*.

Si la première solution en matière d'anonymisation consiste à masquer les informations à anonymiser, ce choix appauvrit la qualité des informations initialement présentes et induit une perte d'information concernant la catégorie sémantique de l'information anonymisée. En ce qui concerne les dates, un écart temporel entre deux dates peut se révéler important au point de vue clinique, et la distinction de plusieurs intervenants *(patient, chirurgien, membres de la famille)* dans un compte rendu hospitalier paraît essentielle à préserver. L'utilisation des balises XML typantes sera donc privilégiée sur l'effacement complet de l'information anonymisée.

Utilisation de pseudonymes

Une solution alternative qui permet de conserver la sémantique des informations anonymisées consiste à remplacer les informations du document par des informations vraisemblables de même catégorie *(un prénom par un autre, une date par une autre, etc.)*, en veillant à respecter deux points : (i) toujours remplacer un prénom par le

même prénom dans la suite du document, et (ii) conserver les écarts temporels réels entre dates fictives. On parle de « *pseudonymisation* ».

Cette méthode repose donc sur : (i) la propagation des identifiants utilisés pour anonymiser des informations nominatives *(propager sur l'ensemble du document l'identifiant numérique ou le pseudonyme associé à un même nom)* et (ii) le remplacement de toutes les dates présentes dans le document en conservant l'écart temporel entre chacune d'entre elles. La propagation des identifiants permet de conserver l'information relative à la personne dont on parle, donc de discriminer les différents intervenants mentionnés dans le document *(par exemple, le patient sera toujours présenté sous le nom « Jean Dupont », le chirurgien l'ayant opéré sous le nom « Paul Martin », etc.)*.

Une version simplifiée de cette approche a été utilisée par [Pestian et al., 2012] pour constituer un corpus anonymisé de notes de suicide, corpus qui a été utilisé lors du challenge i2b2/VA 2010. Dans ce corpus, les prénoms masculins ont tous été remplacés par *John*, les prénoms féminins par *Jane*, les adresses postales par celle du CCHMC, et les dates par une date aléatoirement tirée avec conservation de l'année d'origine. Ce type de traitement basique suffit pour le corpus considéré, les notes que les gens rédigent avant un suicide contenant peu de données personnelles *(le destinataire de ces notes connait le nom de famille, la personne signe donc de son prénom et ne mentionne ses proches que par leur prénom)* et un nombre réduit d'occurrences de chaque type *(peu de prénoms, peu de dates, etc.)*.

Les documents que nous avons à traiter étant des documents cliniques et non des notes de suicide, ils intègrent davantage d'informations personnelles ; les traitements que nous devons appliquer devront donc prendre en compte une complexité plus grande. Cependant, le principe d'utilisation de pseudonymes retient notre attention, et c'est ce principe que nous comptons appliquer sur nos corpus.

Puisque le masquage des informations au moyen de balises ou d'une suite de caractère appauvrit la richesse du texte pour certains types d'information, mais parce qu'il peut être complexe de gérer la propagation des informations dans le document en termes de pseudonymisation, il peut être fait appel à une combinaison des deux précédentes méthodes : (i) utiliser des balises XML typantes avec un identifiant numérique unique pour chaque occurrence *(permet de faire la distinction entre le patient et le chirurgien)* pour les informations nominales, et (ii) utiliser des dates fictives tout en conservant les écarts temporels.

k-anonymat et l-diversité

Une dernière solution consiste à généraliser les informations au moyen d'hyperonymes. Elle est mise en œuvre par le modèle du k-anonymat et de son extension, la l-diversité. Le modèle du k-anonymat [Sweeney, 2002] s'inspire des conséquences issues des expériences de réidentification de la population américaine à partir de triplets d'information [Sweeney, 1996, Sweeney, 2000] (voir section 1.3.3).

Partant de ces expériences, le modèle du k-anonymat consiste à généraliser les informations personnelles sous des classes plus génériques *(une tranche d'âges au lieu d'un âge, le nom d'une région au lieu du nom d'une ville, le nom d'une ville au lieu de l'adresse exacte, etc.)* de telle sorte qu'au terme du processus de généralisation, il y ait au moins k individus dans chaque groupe d'informations. L'utilisateur spécifie donc le nombre d'individus qu'il souhaite obtenir au final dans chaque groupe en fixant la valeur de k. Ainsi, à partir des triplets d'information de base (exemple 1.a),

le modèle appliqué pour générer un groupe d'anonymat où $k = 3$ produira la sortie [13] de l'exemple 1.b.

(1)

a. <30 ans, Paris, Rhume>
<32 ans, Versailles, Grippe>
<29 ans, Créteil, Gastroentérite>

b. <[29-32] ans, Ile-de-France, Rhume>
<[29-32] ans, Ile-de-France, Grippe>
<[29-32] ans, Ile-de-France, Gastroentérite>

Un inconvénient majeur du modèle k-anonymat mentionné par les participants du projet DEMOTIS [14] repose sur le fait que, bien que l'utilisateur ait fixé le nombre de groupes qu'il souhaite produire au moyen de la valeur renseignée pour k, il n'est pas impossible que dans un groupe d'individus partageant tous « *la même information critique (e.g., la même pathologie)* », le modèle du « *k-anonymat ne fournira aucune protection supplémentaire* ». [15]

C'est pour pallier ce problème qu'a été proposée la l-diversité [Machanavajjhala et al., 2006]. Dérivée du k-anonymat, la l-diversité impose que dans chaque groupe de k individus, il y ait au moins l valeurs sensibles différentes. Dans l'exemple 1.b, on produit un groupe composé de trois individus *(3-anonymes)* mais également de trois diversités *(3-diverses)*. Dans l'exemple 2, on produit un groupe de trois individus *(3-anonymes)* composé d'une seule diversité *(1-diverse)*.

(2) <[29-32] ans, Ile-de-France, Rhume>
<[29-32] ans, Ile-de-France, Rhume>
<[29-32] ans, Ile-de-France, Rhume>

Nous estimons que ce mode de représentation avec conservation des informations d'origine sous une forme plus générique *(hyperonymes et intervalles de valeurs numériques)* présente cependant un inconvénient majeur pour les méthodes à base d'apprentissage statistique. En effet, l'apprentissage statistique se fonde sur un lien entre informations à anonymiser et contexte d'apparition (voir section 2.3). Avec ce mode de représentation, les outils d'apprentissage devront donc construire des modèles pour chaque hyperonyme et chaque intervalle de valeurs, conduisant à une baisse globale de l'efficacité du système ainsi construit.

1.4 La catégorisation des informations à anonymiser

On distingue plusieurs catégories d'information qui sont suceptibles de permettre l'identification d'une personne et qui doivent donc faire l'objet d'une anonymisation. L'appartenance d'un type d'information à l'une ou l'autre de ces catégories conditionne fortement les choix logiciels et algorithmiques qui devront être mis en œuvre pour procéder à son anonymisation.

13. Exemples tirés du site http://consultation.demotis.org/glossaire/k-anonymat relatif au projet DEMOTIS.

14. http://www.demotis.org, Définir, Evaluer et MOdéliser les Technologies de l'Information de la Santé. Projet SopinSpace, INRIA, Cecoji-CNRS. Financement ANR-STIC 2009-2012. L'objectif de ce projet visait la conception de systèmes d'information personnelle de santé selon une approche réunissant informaticiens et juristes.

15. http://consultation.demotis.org/doc/3-moduler-lacces-aux-donnees-pour-buts\-epidemiologiques-ou-devaluation-clinique-au-dela-de-linvo

1.4.1 Informations nominatives et numériques

Principe. La première catégorie regroupe tous les types d'informations qui permettent d'identifier une personne. Sont ainsi regroupées sous cet intitulé les données nominatives telles que les noms, les prénoms, les noms de lieux *(adresses postales, villes, etc.)*. On regroupe également dans cette catégorie toutes les informations numériques qui peuvent être associées à une personne *(date de naissance, d'admission, d'opération, de sortie, numéro de téléphone, numéro de télécopie, numéro de sécurité sociale, de mutuelle, code postal, etc.)*.

Le traitement des informations relevant de cette première catégorie correspond à la définition du terme anglais *anonymization*. Un repérage d'entités nommées relevant de catégories prédéfinies est appliqué avant de procéder à l'anonymisation des entités ainsi identifiées.

Si les données nominatives et numériques relatives au patient *(nom, prénom et dates de naissance)* semblent constituer les informations les plus sensibles qu'il convient de traiter, il est légitime de s'interroger sur l'existence d'une tolérance quant à un manque d'anonymisation de certaines catégories d'information *(le nom d'une ville ou d'un hôpital)*. Citant le paragraphe 164.514 du HIPAA, [Uzuner et al., 2008] précisent ainsi que le risque de réidentification doit être le plus faible possible sans nécessairement être nul.

Ambiguïté. D'autre part, [Uzuner et al., 2007] relèvent que l'anonymisation recèle un certain degré d'ambiguïté qu'il n'est pas simple de traiter. On peut ainsi relever plusieurs cas d'ambiguïtés :

- Une ambiguïté de catégorisation pour les entités homographes *(deux entités différentes partageant une même forme graphique : « Paris » est soit un prénom, soit le nom d'une ville)*. Dans ce cas, l'anonymisation de l'entité ne fait aucun doute, en revanche se pose la question du choix de la catégorie à attribuer à l'étiquette, la désambiguïsation s'effectuant en contexte ;
- Une ambiguïté liée à une entité relevant d'une catégorie mais ne devant être anonymisée dans certains cas uniquement *(c'est notamment le cas des noms de maladie qui intégrent un nom propre, ce nom ne devant pas être anonymisé : « maladie de Parkinson », « maladie de Wernig Hoffmann », « fièvre de Lassa », etc.)* ;
- Et une ambiguïté liée à la signification des acronymes : une distinction doit être faite entre les initiales d'une personne *(« PO »)* et un acronyme utilisé en médecine *(« po »,* per os*, par voie orale)*.

La législation française

En France, la Commission Nationale Informatique et Libertés (CNIL) fournit une fiche préconisant une méthode d'anonymisation. [16] Cette fiche se place dans un contexte d'étude épidémiologique du virus du SIDA, et définit des recommandations de respect de l'anonymat lors de la collecte des données. La CNIL dresse ainsi brèvement un ensemble de types de données *(les données directement rattachées au patient — nom, prénom, date de naissance, etc. — et celles qui le sont indirectement — matricule, adresse, numéro de téléphone, élément biométrique, adresse IP, traces*

16. http://www.cnil.fr/en-savoir-plus/fiches-pratiques/fiche/article/letat-des-lieux-en-matiere-de-procedes-danonymisation

de données de connexion, etc.) devant faire l'objet d'une anonymisation, soit au moment du recueil des données, soit *a posteriori*, avant de les enregistrer dans une base de données.

Toujours dans un contexte d'anonymisation effectuée à l'occasion d'un recueil des données, la CNIL a dressé dans cette fiche une liste de préconisations que nous reproduisons ci-dessous :

- Ne collecter les données qu'au niveau de finesse strictement nécessaire ;
- Répartir les données dans des fichiers ou systèmes informatiques distincts pour éviter toute réidentification par le biais du croisement des informations ;
- Cloisonner le recueil et la saisie des données entre plusieurs personnes et organismes ;
- Ne pas utiliser d'outil d'interrogation des données permettant le croisement des données ;
- Interdire certains croisements ;
- Ne pas afficher les résultats d'une requête si le nombre de documents correspondant à la requête est inférieur à dix.

Il résulte du contenu de cette fiche que l'anonymisation n'est prévue par la CNIL que dans le cadre d'une collecte de données, et non dans celui de l'utilisation de comptes rendus cliniques. Il est par ailleurs intéressant de constater que ce type de document enfreint certaines des préconisations listées ci-dessus : (i) les données ont été saisies par la même personne *(un membre de l'équipe chirurgicale qui a opéré le patient)*, (ii) les données sont rassemblées dans le même document *(données nominatives et numériques, données cliniques)*, et (iii) les données présentes dans ces documents correspondent à un impératif clinique que l'on peut considérer comme relevant du « niveau de finesse strictement nécessaire ».

Cependant, en préconisant l'utilisation d'une clé de hachage pour anonymiser les données, la CNIL précise quels sont les points importants qui l'intéressent en matière d'anonymisation :

- Le caractère irréversible de l'anonymisation, c.-à-d. l'impossibilité de réidentification du patient ;
- Le très faible taux de « collisions » en cas d'utilisation d'une clé de hachage, c.-à-d. le risque qu'une même clé soit attribuée à deux individus distincts.

Par extrapolation, dans le cadre de comptes rendus cliniques, on cherchera donc à anonymiser toutes les informations qui permettent, directement ou indirectement, d'identifier un individu, avec pour objectif final l'impossibilité de réidentifier l'individu sur la base des informations qui auront été conservées en clair.

La législation américaine

En 1996 aux États-Unis, le Congrès a voté la loi dite « *Health Insurance Portability and Accountability Act* » (HIPAA) dans le domaine de la santé et de l'assurance maladie. Dans le cadre des règles de respect de la vie privée *(Privacy Rules)* de cette loi, les informations personnelles devant faire l'objet d'une anonymisation pour qu'une redistribution des données soit envisagées ont été rassemblées dans une liste de dix-huit identifiants. [17]

17. Les dix-huit identifiants du HIPAA ainsi que la définition de ce qui constitue une information à anonymiser sont rappelés à l'adresse suivante : http://www.research.ucsf.edu/chr/HIPAA/

Ces informations sont qualifiées de « *Protected Health Information* » (PHI). Elles renvoient à n'importe quel type d'information présent dans les documents cliniques qui permet d'identifier un individu et qui a été créé, utilisé ou mis à disposition dans le cadre d'un service de soin. La loi autorise cependant les chercheurs a accéder aux données privées lorsque cela s'avère nécessaire pour la recherche effectuée.

[Deléger et al., 2013] rappellent que la loi HIPAA implique deux contraintes particulières : (i) la suppression de l'ensemble des PHI dans les documents, et (ii) l'obtention du consentement des participants. Cependant, puisque la demande de consentement réduit le taux de participation, l'anonymisation complète des documents demeure la meilleure solution [ibid.]. D'autre part, [El Emam et al., 2009] considèrent que la demande explicite du consentement auprès des patients peut conduire à un biais dans la sélection et le recrutement des patients. Les auteurs ont également considéré que la meilleure solution passait par le développement d'un outil d'anonymisation, fondé sur le critère k-anonymat (voir section 1.3.5).

Enfin, [McGraw, 2013] souligne que le HIPAA est critiqué sur plusieurs aspects, parmi lesquels : (i) le manque de méthodologies pour anonymiser les données *(soit la méthode statistique dans laquelle le statisticien atteste que les données traitées présentent un faible risque de réidentification, soit la méthode consistant à éliminer les 18 identifiants)*, (ii) l'absence de réponse juridique aux tentatives de réidentifications non autorisées *(mais il faut préalablement définir ce que recouvre la notion de « réidentification »)*, et (iii) la restriction d'utilisation des données anonymisées à certains cas précis *(par ex., pour des études marketing en pharmacologie)*. Ces critiques conduisent, selon l'auteur, à une baisse de la confiance placée dans ce standard.

Identifiants. La liste des dix-huit identifiants prévus par cette loi est la suivante :

1. Tous les noms et prénoms ;
2. Toute subdivision géographique plus petite qu'un État *(adresse postale, ville, région, code postal,* [18] *et autres codes géographiques)* ;
3. Tout élément de date (à l'exception des années) si la date concerne directement le patient *(date de naissance, date d'entrée, date de sortie, date de décès)*, et les âges strictement supérieurs à 89 ans. [19] Une catégorie unique regroupant tous les âges supérieurs à cette valeur peut être utilisée ;
4. Les numéros de téléphone dans leur intégralité ;
5. Les numéros de télécopie dans leur intégralité ;
6. Les adresses de messagerie électronique ;
7. Les numéros de Sécurité Sociale ;
8. Les références de dossiers médicaux ;
9. Les numéros de mutuelle de santé ;
10. Les numéros de compte ;

`chrHIPAAphi.asp` ; d'autre part, les règles de sécurité et de vie privée du HIPAA sont présentées à l'adresse suivante : http://www.hhs.gov/ocr/privacy/

18. Aux États-Unis, une exception s'applique aux trois premiers chiffres des codes postaux. Si l'unité géographique issue du regroupement de tous les codes postaux partageant les trois premiers chiffres est peuplée de plus de 20 000 habitants, les trois premiers chiffres sont alors conservés à l'identique ; ils sont remplacés par « 000 » le cas échéant.

19. Auquel cas tous les éléments de la date de naissance, y compris l'année, doivent être anonymisés.

11. Les numéros de permis de conduire [20] ;
12. Les identifiants du véhicule *(marque, type, couleur, etc.)* y compris le numéro d'immatriculation ;
13. Les références d'appareillage médical *(défibrillateur cardiaque, prothèses diverses, etc.)* et leur numéro de série ;
14. Les adresses Internet (URL) ;
15. Les adresses IP (adresse électronique identifiant un ordinateur sur le réseau) ;
16. Les identifiants biométriques *(empreintes digitales, empreintes vocales, etc.)* ;
17. Les photographies du visage et n'importe quelle image similaire ;
18. Et de manière plus générale, n'importe quel autre identifiant ou caractéristique unique permettant l'identification du patient *(cicatrice, tatouage, etc.)*.

Informations hors HIPAA. Cette législation précise par ailleurs que l'identifiant unique attribué par celui qui anonymise les données ne relève pas de cette dernière catégorie. Ce nouvel identifiant peut donc être utilisé dans les documents ainsi anonymisés dans le cadre de futures recherches, par exemple pour rassembler les différents documents se rapportant au même patient.

D'autre part, l'anonymisateur doit veiller au fait que parmi les informations non affectées par le processus d'anonymisation, celles-ci ne permettent pas la réidentification du patient. Autrement dit, les informations ne relevant pas des dix-huit catégories précédentes mais qui se révèlent néanmoins identifiantes (par combinaison avec des connaissances externes par exemple) doivent faire l'objet d'une anonymisation également.

Aussi exhaustive et précise que soit cette liste d'identifiants, il n'existe aucun outil permettant de traiter l'ensemble des catégories. Pour pallier ce manque, [Deléger et al., 2013] ont travaillé sur la comparaison de deux outils existants à base d'apprentissage *(Mallet et MIST)* pour traiter douze de ces catégories *(noms et prénoms, adresses géographiques, dates, âges, téléphones et télécopies, adresses de messagerie électronique, numéro de Sécurité Sociale, numéro de permis de conduire, initiales, institutions, adresses IP, et une catégorie « autre »*, voir section 2.4*)*.

Alors que la majorité des équipes travaillant en anonymisation se fonde sur les catégories du HIPAA, [Malin et al., 2011] rappellent qu'il existe une alternative au HIPAA qui se révèle cependant méconnue, reposant sur le « standard statistique ». Les auteurs regrettent par ailleurs le manque de publication de méthodologie sur cette alternative, ce qui ne permet pas de répandre son utilisation.

1.4.2 Informations préjudiciables

Principe. La seconde catégorie d'information concerne tous les éléments qui peuvent porter préjudice à une personne. Ce type d'information n'est présent que dans certains genres de corpus, tels les corpus oraux dans lesquels sont présents des échanges entre plusieurs locuteurs, parmi lesquels certains peuvent tenir des propos discourtois à l'encontre des autres personnes de l'assistance.

20. Aux États-Unis, le permis de conduire sert de carte d'identité. Par extension, n'importe quel numéro d'identité est inclus dans cette catégorie (numéro de passeport par exemple).

Les informations qui peuvent porter préjudice envers d'autres personnes sont plus complexes à traiter : elles peuvent être de nature injurieuse ou plus subtiles *(reproches, surnoms, jeux de mots)*. Il apparaît cependant que dans les corpus de documents cliniques, ce type d'information n'a aucune chance d'être présent. On n'imagine en effet assez difficilement un professionnel de la santé faire référence à son patient dans un document écrit au moyen de qualificatifs négatifs ou péjoratifs ; tout au plus une pointe d'ironie percera t-elle par le biais d'une antiphrase *(notamment en qualifiant un patient de « sympathique » alors qu'il s'agit d'un patient particulièrement difficile)*.

Les jeux de mots et surnoms relatifs à des personnes, parce qu'ils sont complexes et intrinsèquement liés au discours dans lesquels ils s'insèrent ou à des références culturelles dans lesquelles évoluent les personnes concernées, posent de nombreuses questions à l'anonymisateur. Faut-il effectuer une anonymisation sur le jeu de mot produit ? Dans quelle mesure le jeu de mots ou le surnom ainsi créé est-il identifiable ?

Témoignage. [Reffay et Teutsch, 2007] rapportent l'exemple suivant provenant d'un corpus d'interactions entre apprenants du français et dans lequel une intervenante nommée Bianca se présente à un autre intervenant : « *en la Colombie mes amies m'appelent contradiction parce que Bianca signifie blanc et je suis un peu marron, tres marron.* ». L'intervenante est affublée d'un surnom dans son pays natal, surnom qui correspond à la relation qui existe entre la signification de son prénom et sa couleur de peau. Dans une procédure d'anonymisation, le prénom comme le surnom devraient tous deux être masqués. La traduction du prénom reste cependant apparente et potentiellement réidentifiante...

1.4.3 Combinaison d'informations

Présentation

Cette dernière catégorie se révèle particulièrement complexe. Elle a émergé ces dernières années sous le coup des avancées technologiques offertes par les outils informatiques. [Baude, 2006] rappelle ainsi que les procédures permettant de réidentifier des personnes ont bénéficié des récents apports technologiques tels que les facilités offertes en matière de stockage et de diffusion des informations, ainsi que la puissance de calcul accrue pour traiter de gros volumes de données.

Cette catégorie rassemble toutes les informations qui, prisent isolément, n'offrent aucun pouvoir d'identification mais qui, combinées entre elles, permettent de réduire le champ des possibles, voire de remonter jusqu'à une réidentification. Le recoupement d'informations est possible aussi bien à l'intérieur d'un document qu'entre plusieurs documents.

Précisons toutefois qu'à ce niveau, l'identification du patient ne repose plus sur un élément directement présent dans le document clinique — je connais le patient parce que son nom de famille est renseigné dans une phrase —, mais elle repose sur une démarche volontaire de recherche et de recoupement d'informations. À ce niveau, la réidentification suppose une volonté forte de remonter jusqu'au patient, jusqu'à mobiliser les moyens nécessaires à cette entreprise *(programme complexe de traitement des informations, accès aux bases de données, recherche sur internet, etc.)*.

Cas d'étude

Afin de comprendre ce processus, imaginons le dossier médical d'un patient composé de trois documents anonymisés du point de vue des informations nominatives *(noms, prénoms, villes, etc.)* et numériques *(numéro de sécurité sociale, code postal, dates, etc.)*, autrement dit, un dossier uniquement composé d'informations cliniques.

Dans le premier document *(résultats de laboratoire)*, le groupe sanguin du patient est mentionné. Prenons un groupe sanguin peu présent dans la population tel que le groupe sanguin AB- (voir tableau 1.2, source : *Établissement Français du Sang*, 29 avril 2011).

	O	A	B	AB
Rhésus +	36 %	37 %	9 %	3 %
Rhésus -	6 %	7 %	1 %	1 %

TABLE 1.2 – Répartition des groupes sanguins dans la population française en 2011

À l'échelle de la population française (64 612 939 habitants selon le recensement 2010 de l'INSEE), 646 129 français (1 %) seraient donc du groupe AB-.

Dans le deuxième document *(compte rendu opératoire)*, on apprend que le patient a été victime du syndrome de Brugada *(maladie cardiaque correspondant à une fibrillation ventriculaire idiopathique également appelée « syndrome de mort subite nocturne inattendu »)*. D'après le site *Orphanet*,[21] la prévalence du syndrome de Brugada en Europe et aux États-Unis est de 1/10 000. En postulant que ce syndrome affecte une population d'individus indépendamment de son groupe sanguin, la population combinant groupe sanguin AB- et syndrome de Brugada s'élèverait à 64 personnes en France (un cas sur dix-mille parmi 646 129 français).

Dans le dernier document *(lettre de suivi)*, on apprend que le patient est une femme (soit par le biais des pronoms personnels et des règles d'accord linguistique en genre, soit parce qu'étant explicitement mentionné : « *J'ai revu votre patiente, Mme...* »). Le site *Orphanet* nous précise que le syndrome de Brugada affecte davantage les hommes que les femmes, selon un ratio de 8:1. Sur les 64 français combinant syndrome de Brugada et groupe sanguin AB-, seulement sept seraient des femmes. La patiente correspondant à ce dossier serait donc l'une de ces sept françaises...

Toutes ces informations ne sont pas identifiantes en tant que telles. Néanmoins, en recoupant ces données, et par le jeu des éliminations dans une population d'individus, pour peu que l'on dispose des accès aux bases d'informations contenant les indications pertinentes, il est possible de circonscrire un nombre relativement faible de patients potentiels, voire de remonter jusqu'à la personne dont on parle dans le dossier. Pour cette dernière étape, il importe de relier ces précédents recoupements avec des informations nominatives *(par exemple au moyen des réseaux sociaux où les informations personnelles abondent)*.

21. http://www.orpha.net/, Orphanet, « Le portail des maladies rares et des médicaments orphelins » (INSERM, Ministère de la Santé, Union européenne).

Conclusion

Ce processus de traitement des possibilités de recoupement et de rattachement d'informations correspond à la réalité couverte par le terme anglais *de-identification*. Si l'opération de réidentification du patient par le biais des recoupements d'informations distribuées entre documents s'avère complexe, la tâche initiale de définition des informations devant être anonymisées pour éviter cette réidentification se révèle également complexe. Précisons cependant qu'il convient de conserver un juste équilibre entre informations devant être anonymisées *(nominatives et numériques)* et celles, d'ordre clinique, demeurant en clair dans le document.

Les traces produites par le système *(plateformes de téléformation à distance)*, utiles pour améliorer le système ou pour interroger et caractériser le système, doivent faire l'objet d'un traitement particulier. D'autre part, les identifiants numériques créés par l'anonymiseur n'ont pas à être anonymisés !

1.5 L'anonymisation dans les projets de recherche scientifique

1.5.1 Processus d'anonymisation

Cependant, une problématique forte se pose quant aux possibilités de développer de tels outils. En effet, comment peut-on envisager le développement d'outils d'anonymisation sans travailler sur des données réelles, seul moyen de garantir la robustesse du système ? Deux solutions existent pour pallier ce problème.

La première solution consiste à remplacer les données existantes qui sont identifiantes *(noms, prénoms, lieux)* par des données vraisemblables *(noms et prénoms génériques, en attribuant le même prénom générique à chaque occurrence d'un patient)*. C'est notamment l'approche qui a été suivie par [Uzuner et al., 2007] dans le cadre du challenge international i2b2 2006 dont la thématique portait sur l'anonymisation automatique de comptes rendus cliniques.

La seconde solution que nous avons suivie dans la mise au point de notre chaîne d'anonymisation consiste à anonymiser les données les plus sensibles au sein de l'hôpital, au moyen d'une recherche à l'identique des noms et prénoms présents dans le système d'information patient (SIP) de l'hôpital. Cette première étape permet ainsi de sortir les données pour développer les outils d'anonymisation en dehors du parcours de soins classique.

1.5.2 Intérêt des outils d'anonymisation

Malgré l'intérêt porté à ces différents corpus, une contrainte forte pèse sur la disponibilité des corpus porteurs d'informations personnelles en dehors du parcours de soin : ces corpus doivent faire l'objet d'une anonymisation [Zweigenbaum, 2008]. Alors que les corpus américains utilisés dans le cadre des campagnes d'évaluation suivent les principes du HIPAA pour procéder aux anonymisations requises, en France, il n'existe aucune liste précise d'identifiants à traiter. La CNIL fournit cependant une liste de recommandations à prendre en compte, en particulier lors du recueil des données (voir section 1.4.1).

Projets de recherche

On retrouve cet impératif d'anonymisation dans plusieurs projets de recherche actuels dont le matériau de départ repose sur des corpus de comptes rendus cliniques.

Le projet Akenaton. Il en est ainsi du projet Akenaton [22] dans lequel s'inscrit majoritairement ce travail de recherche. Ce projet concerne l'extraction d'informations médicales depuis des comptes rendus cliniques rédigés en texte libre, à propos de patients équipés d'un défibrillateur cardiaque (voir section 4.3.1).

Le projet ALADIN-DTH. Un autre projet — le projet ALADIN-DTH [23] — vise le développement d'un outil de détection automatique des infections nosocomiales, à partir de documents médicaux rédigés en langage naturel. La mise au point de cet outil a demandé une étape d'anonymisation préalable. Pour ce faire, un outil d'aide à l'anonymisation du contenu des comptes rendus cliniques a été développé par Pierre Marchal à l'intention des médecins hygiénistes des établissements partenaires du projet [Gicquel et al., 2012] (voir section 2.2). Il repose sur une première étape de pré-anonymisation automatique *(règles et projection de lexiques)* puis sur une deuxième étape de correction manuelle pour laquelle il est possible de propager l'étiquette choisie par l'utilisateur humain pour une entité donnée à l'ensemble du document. Cet outil, installé sur un ordinateur portable, a été utilisé à l'intérieur de chaque CHU partenaire avant la mise à disposition des données aux membres du consortium.

Le projet Accordys. Un dernier exemple de besoin d'outils d'anonymisation concerne le projet Accordys [24] qui vise la recherche de cas similaires dans des documents cliniques, dans un contexte de dysmorphologie fœtale. À l'image des projets précédents, la mise au point des outils permettant cette recherche suppose de travailler sur des données anonymisées. La chaîne d'anonymisation produite à l'occasion du projet Akenaton sera réutilisée et adaptée aux nouveaux contenus de ce projet.

Pour conclure, il importe de distinguer deux étapes majeures dans les projets de recherche : (i) une première étape de mise au point des outils de traitement de l'information médicale par les partenaires du projet, et (ii) l'application de l'outil développé, sur des données réelles. Puisque la première étape est généralement accomplie par les partenaires du projet, en dehors de l'hôpital, l'anonymisation des données est obligatoire. En revanche, la deuxième étape est réalisée à l'intérieur de l'hôpital qui aura fourni les données d'entraînement. À ce titre, l'anonymisation des données n'est donc plus utile.

On retiendra au final que dans le cadre des projets de recherche, l'anonymisation des informations personnelles contenues dans les données cliniques n'a pour seul but que de permettre la sortie de ces données hors de l'hôpital, afin de mettre au point les outils de traitement de l'information.

22. http://resmed.univ-rennes1.fr/AKENATON/, Automated Knowledge Extraction from medical records iN Association with a Telecardiology Observation Network, financement ANR TecSan 2007/2010.

23. http://www.aladin-project.eu/, Assistant de Lutte Automatisée et de Détection des Infections Nosocomiales à partir de Documents Textuels Hospitaliers, financement ANR TecSan 2009/2011.

24. http://ics.upmc.fr/node/138, Agrégation de Contenus et de COnnaissances pour Raisonner à partir de cas de DYSmorphologie fœtale, financement ANR ConInt 2012/2014.

Intérêt de la communauté

En dehors de tout projet de recherche, on retrouve des témoignages de l'intérêt porté par la communauté scientifique à ce sujet au travers de trois types de manifestation : (i) les communications orales lors de conférences,[25] (ii) les articles parus en revue, et (iii) les campagnes d'évaluation sur ce sujet.

Les communications et publications. À titre d'exemple, nous avons relevé par année le nombre d'articles de conférences et de revues indexés par MEDLINE, correspondant à l'une des quatre requêtes suivantes dans PubMed : « *de-identification* » (86 articles), « *de-identification text* » (22 articles), « *anonymization text* » (4 articles) et « *text scrubbing* » (4 articles).[26] Nous donnons dans le graphique 1.1 les résultats de ce décompte effectué le 28 décembre 2012,[27] après avoir éliminé les doublons liés aux requêtes. Nous notons une évolution notable des publications sur le sujet, avec une franche progression après 2006 *(année où s'est tenu le défi i2b2 sur la thématique de l'anonymisation)*. Cette évolution traduit bien évidemment l'intérêt porté par la communauté, mais elle reflète également les recherches effectuées pour répondre aux impératifs d'anonymisation édictés par les différentes lois.

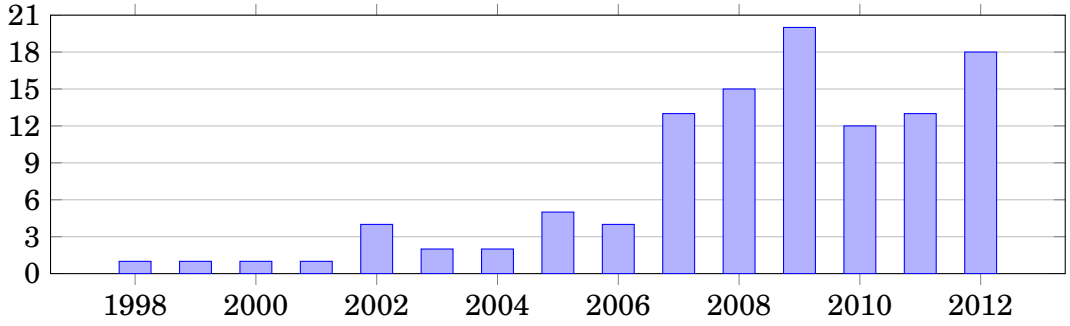


FIGURE 1.1 – Nombre d'articles indexés par année dans MEDLINE sur la thématique de l'anonymisation

Les campagnes d'évaluation. Hormis les publications, une expression forte de l'intérêt suscité par cette problématique réside dans l'organisation de campagnes d'évaluation ou la participation à ces campagnes.

Présentation. En 2006, l'institut américain i2b2[28] a organisé un challenge sur le thème de l'anonymisation automatique [Uzuner et al., 2006, Uzuner et al., 2007].

Huit catégories d'information ont été définies comme devant être anonymisées : (i) les noms des patients et membres de la famille *(prénom, nom, mais pas les titres : « Mr., Ms »)*, (ii) les noms des médecins et autres professionnels de la santé *(noms,*

25. Symposium de l'AMIA, MEDINFO, etc.

26. Les revues suivantes rassemblent les articles renvoyés par cette requête : *BMC Bioinformatics* (2 articles), *BMC Med Inform Decis Mak* (4 articles), *BMC Med Res Methodol* (2 articles), *Int J Med Inform* (3 articles), *J Am Med Inform Assoc* (15 articles), *J Biomed Semantics* (1 article), *Med Care* (1 article), *Methods Inf Med* (2 articles), *Yearb Med Inform* (1 article). Le symposium annuel de l'AMIA rassemble 10 articles.

27. À cette date, tous les articles de l'année 2012 sont présents dans la base.

28. https://www.i2b2.org/, Informatics for Integrating Biology & the Bedside, Boston, MA.

prénoms, initiales, mais pas les titres : « Dr., Pr. »), (iii) les noms des hôpitaux et maisons de repos, (iv) les identifiants alpha-numériques *(comptes rendus cliniques, patients, docteurs, hôpitaux)*, (v) les dates, y compris les années, (vi) les lieux *(villes, rues, noms des rues, codes postaux, noms de bâtiment, numéro)*, (vii) les numéros de téléphone, et $(viii)$ les âges, y compris ceux inférieurs à 90 ans.

Le corpus se composait de 889 comptes rendus cliniques, à raison de 669 documents pour l'apprentissage et 220 pour le test. Trois annotateurs humains ont remplacé les données identifiantes d'origine par des données vraisemblables *(données fictives)*. Sept équipes ont participé à cette première édition.

Méthodes. Dans le cadre du challenge i2b2 2006, la majorité des équipes (cinq sur sept, voire six [29]) a eu recourt aux techniques par apprentissage statistique, soit de manière exclusive, soit en combinaison avec des règles. Trois formalismes ont ainsi été utilisés : des CRF *(via CRF++)*, des SVM *(au moyen de SVMlight ou de LibSVM)* et des arbres de décision *(algorithme C4.5)*. Certains de ces formalismes ont été intégrés dans des plates-formes d'extraction d'information, telles que Carafe ou GATE.

Un grand nombre d'équipes a ainsi envisagé le problème comme relevant d'une tâche de reconnaissance des entités nommées dont l'issue consisterait à masquer les informations ainsi identifiées [Gardner et Xiong, 2009, Szarvas et al., 2007, Wellner et al., 2007]. D'autres équipes ont perçu le challenge comme une tâche de classification entre éléments à anonymiser ou non [Guo et al., 2006].

Évaluation et discussion. Les résultats [30] obtenus par les participants (voir tableau 1.3 ou figure 1.2) permettent de dresser les conclusions suivantes, la comparaison ayant été rendue possible par l'application des différents systèmes sur le même jeu de données. Ces éléments de conclusion restent largement dépendants du corpus testé et des catégories couvertes dans le cadre du challenge.

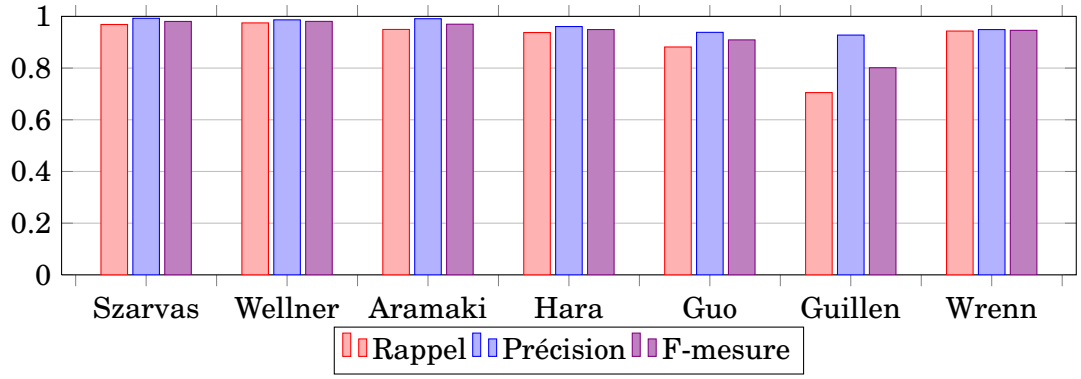


FIGURE 1.2 – Résultats des participants du challenge i2b2 2006 classés par méthode

Les méthodes à base de règles sont moins robustes que les méthodes par apprentissage statistique (F=0,8013). Elles ne permettent pas d'obtenir un rappel élevé

29. L'équipe Wrenn n'a pas fourni de description de son système aux organisateurs du challenge : « *Wrenn et al., the engineers of the seventh system, participated in the challenge but provided us with no system description* » [Uzuner et al., 2007]. Si les résultats obtenus par cette équipe penchent en faveur d'un apprentissage statistique, rien ne permet de l'assurer avec certitude ni d'identifier le formalisme utilisé.

30. Les définitions et formules des mesures d'évaluations utilisées *(rappel, précision, F-mesure, etc.)* sont données dans le chapitre 3.

Équipe	Szarvas	Wellner	Aramaki	Hara	Guo	Guillen	Wrenn
Méthode	Arbres de décision	CRF	CRF	Règles + SVM	SVM	Règles	Inconnue
Rappel	0,9686	0,9747	0,9494	0,9375	0,8816	0,7052	0,9432
Précision	0,9926	0,9865	0,9909	0,9607	0,9382	0,9276	0,9489
F-mesure	0,9804	0,9806	0,9697	0,9490	0,9090	0,8013	0,9460

TABLE 1.3 – Résultats des participants du challenge i2b2 2006 classés par méthode

même si leurs performances qualitatives s'avèrent bonnes (précision équivalente à celle obtenue par les autres méthodes).

Toutes les méthodes intégrant de l'apprentissage ont permis l'obtention de résultats élevés, avec des F-mesures supérieures à 0,909 et une précision toujours supérieure au rappel. Les résultats se révèlent proches entre équipes, quel que soit le formalisme retenu, même si les arbres de décision et les CRF semblent avoir mieux réussi que les SVM.

La combinaison de systèmes à base de règles et d'apprentissage statistique permet, soit d'aider à la sélection des caractéristiques pour la construction du modèle, soit d'effectuer des post-traitements. L'utilisation de règles en complément du SVM semble préférable (F=0,9490) à une utilisation simple du SVM (0,9090), améliorant tant le rappel que la précision. Nous employons le verbe « sembler » dans cette comparaison car il serait vain de considérer que ce facteur permet à lui seul d'expliquer un gain global de 4 points. S'agissant de deux systèmes différents, produits par deux équipes distinctes avec l'utilisation de caractéristiques *a priori* différentes pour construire les modèles, il n'est guère possible de connaître avec précision le gain apporté par l'ajout des règles.

1.6 Le couteau suisse de l'anonymisation existe t-il ?

Sachant qu'un anonymiseur se doit de masquer toutes les informations qui permettent la réidentification du patient, et parce qu'il est obligatoire d'anonymiser les documents cliniques avant de les communiquer hors de l'hôpital (quel que soit le pays), il est légitime de s'interroger sur la disponibilité d'un outil d'anonymisation universel. S'il n'existe à l'heure actuelle aucun outil universel, un tel outil devrait être capable de gérer les aspects suivants :

1. Le multilinguisme : est-il possible de traiter les documents cliniques rédigés dans plusieurs langues au moyen d'un seul outil ? Limitons cet aspect aux langues utilisant l'alphabet latin *(l'anglais, le français, le suédois, etc.)* pour évacuer les différences liées aux alphabets.
2. Le multi-culturalisme : est-il possible de disposer d'un outil traitant les documents provenant de plusieurs pays qui partagent la même langue *(par exemple dans la sphère de la francophonie)* ? À un niveau plus restreint, est-il envisageable de traiter les documents provenant de plusieurs centres hospitaliers d'un même pays, voire de plusieurs services d'une même spécialité ?
3. Et le multi-disciplinaire : un seul outil est-il en mesure de traiter les documents issus de différents services hospitaliers *(cardiologie, radiologie, laboratoire d'analyse, etc.)* ?

Si la perspective d'un anonymiseur universel et générique est séduisante, elle se révèle malheureusement impossible à mettre en place pour les raisons suivantes.

1.6.1 Un anonymiseur multilingue et multi-culturel

Comme nous l'avons vu en introduction, les langues naturelles sont difficilement formalisables, du fait de l'existence d'un certain nombre d'exceptions à chaque règle *(par exemple, la règle de formation du pluriel par ajout de la désinence « -s » est enfreinte par les mots invariables — une souris, des souris —, les changements de désinences au singulier et au pluriel — un cheval, des chevaux — et les formes différentes d'un même concept au singulier et au pluriel — un œil, des yeux).*

En matière de méthodes symboliques (voir section 2.2), la production des règles formalisant les connaissances d'expert est généralement coûteuse en temps pour un résultat de qualité, mais peu générique. Ce constat implique des temps de développement élevés pour chaque langue.

En matière de méthodes par apprentissage statistique (voir section 2.3), un corpus annoté dans chaque langue sera ainsi nécessaire — processus long et coûteux —, et pour que l'efficacité du système soit maximale, plusieurs modèles devront être produits, *a minima* un modèle par langue, ce qui réduit l'intérêt d'un système universel.

Dans le cadre de ce travail, nous avons d'abord envisagé d'adapter au français un système existant, librement disponible et modifiable : « De-ID » [Neamatullah, 2006, Neamatullah et al., 2008] (voir section 5.3). Cette adaptation au français passe par trois étapes [Grouin et al., 2009b] : (i) l'utilisation de listes dédiées à la langue et à la culture envisagée *(noms, prénoms, et villes utilisés en France)*, (ii) la traduction des déclencheurs, et (iii) l'adaptation des règles, initialement écrites pour l'anglais, au français. Si nous n'avons pas réussi à achever l'adaptation de l'outil pour le français, une équipe de l'Université de Stockholm a pu mener l'adaptation de l'outil à son terme pour le suédois, produisant la version « Deid-Swe ». L'aboutissement positif pour le suédois semble principalement lié à la proximité des deux langues *(deux langues germaniques partageant des structures syntaxiques)*. La similarité de structure des comptes rendus dans les deux langues[31] constitue un deuxième facteur de réussite [Velupillai et al., 2009]. Ce critère peut néanmoins s'appliquer entre différents langues.

À l'occasion de notre tentative d'adaptation de l'outil « De-ID » au français, nous avons par ailleurs pu constater que les règles constituent le point de l'anonymiseur le plus sensible aux contextes multi-culturels. Bien que nous n'ayons pas eu l'occasion de travailler sur des corpus provenant d'hôpitaux francophones situés hors de France, il est aisé de se rendre compte de la difficulté de produire un anonymiseur multi-culturel qui s'appliquerait à des corpus de documents cliniques rédigés en français provenant de différents pays de la francophonie. À titre d'exemple, les différences de format des adresses postales selon que l'adresse est en Belgique (exemple 3.a, le numéro de rue est en fin de ligne, le code postal compte quatre chiffres) ou en France (exemple 3.b, le numéro de rue est en début de ligne, le code postal comprend cinq chiffres) ne permettent pas d'assurer le multi-culturalisme d'un anonymiseur sauf à prendre en compte tous les cas de figure existants.

31. « *The style of the Swedish EPRs and the American English EPRs is very similar when it comes to structure, with notes that describe different sequences in the health care process.* »

(3)

a. Palais Royal
Rue Brederode 16
1000 Bruxelles, Belgique.

b. Centre de Recherche des Cordeliers
15, rue de l'École de Médecine
75006 Paris.

Conclusion. Au vu des expériences d'adaptation au français de l'outil De-ID, il apparaît impossible de prévoir un anonymiseur multilingue. En raison des différences de format de certaines catégories d'informations, telles que les adresses postales, il apparaît également complexe de produire un anonymiseur multi-culturel.

1.6.2 Un anonymiseur multi-disciplinaire

L'impossibilité de disposer d'un anonymiseur universel et générique se manifeste également au travers de la difficile gestion des spécificités relatives à chaque discipline médicale.

Bien qu'un grand nombre de types d'information à traiter soit commune entre les différentes disciplines *(les données nominatives et numériques)*, chaque discipline médicale et chaque corpus possède ses propres caractéristiques qui rendent complexe la production d'un système unique. [Loukides et al., 2010] ont développé leur propre outil pour traiter un corpus sur le génome tandis que [Tu et al., 2010] ont repris un outil existant pour l'adapter aux spécificités de leur corpus de soins primaires.

Des outils doivent ainsi être développés et adaptés pour chaque langue, de manière à traiter le mieux possible les spécificités linguistiques et culturelles de chaque document.

1.7 Synthèse

On distingue deux types de documents médicaux porteurs d'informations personnelles identifiantes : les articles scientifiques, essentiellement des cas d'étude, et les documents cliniques. Les corpus de documents cliniques sont utiles pour mettre au point des systèmes informatiques de traitement de la langue. Cependant, pour que ces documents puissent être utilisés en dehors du parcours de soin classique d'un hôpital, ils doivent avoir fait l'objet d'une anonymisation.

Bien que le terme « anonymisation » ne soit pas relevé dans les dictionnaires que nous avons consultés, chaque locuteur du français l'interprètera comme étant le processus de rendre quelque chose anonyme. L'anglais distingue deux niveaux d'anonymisation en corpus. Un premier niveau, désigné sous le terme « *de-identification* », vise l'impossibilité de relier toute donnée à un patient. Un deuxième niveau, plus précis et désigné sous le terme « *anonymization* », vise à proscrire toute réidentification, y compris par la combinaison d'informations qui, prises isolément, n'offrent aucun pouvoir d'identification. L'objectif final de l'anonymisation et de la désidentification en corpus consiste donc à empêcher au maximum toute réidentification des patients sur la base des informations laissées en clair dans les documents.

Plusieurs techniques existent pour anonymiser les documents cliniques. Préalablement à tout lancement de ce processus, l'anonymisateur doit définir sous quelle forme apparaitront les informations anonymisées : l'effacement par un blanc, le remplacement par une balise typante, l'utilisation de pseudonymes ou d'hyperonymes, tous ces choix doivent être débattus avant tout démarrage de l'anonymisation. Le résultat conditionne la qualité des traitements appliqués ultérieurement sur le corpus anonymisé.

Une catégorisation des informations à anonymiser laisse entrevoir trois classes principales. La première catégorie rassemble toutes les informations nominatives et numériques. Une liste de dix-huit identifiants a été définie dans le cadre de la loi américaine HIPAA, pour représenter les informations à anonymiser dans un corpus médical. La deuxième catégorie renvoie aux informations préjudiciables, cependant absentes des corpus médicaux. Enfin, la dernière catégorie renvoie à la combinaison des informations laissées en clair dans le document, notamment les informations *a priori* non identifiantes.

Du fait des contraintes qui pèsent sur la mise à disposition des corpus, la communauté scientifique témoigne d'un intérêt grandissant pour les outils d'anonymisation, soit dans le cadre des projets de recherche, soit à l'occasion de la participation à des campagnes d'évaluation. Malgré cet intérêt, il semble malheureusement impossible de bénéficier, à l'heure actuelle, d'un outil d'anonymisation universel prêt à l'emploi qui soit multilingue, multi-culturel et multi-disciplinaire.

Première partie

État de l’art

Introduction de la première partie

La première partie de ce manuscrit présente les méthodologies de travail sur lesquelles nous nous fondons pour aborder la problématique de l'anonymisation automatique des documents cliniques, objectif poursuivi dans ce travail de thèse.

❧

Dans un premier temps, nous passons en revue l'état de l'art du domaine en présentant les différentes méthodes utilisées. Ces méthodes relèvent de deux grandes familles, les approches symboliques d'une part, les approches à base d'apprentissage statistique d'autre part. Des expériences d'utilisation détournée de systèmes de reconnaissance d'entités nommées ont également été tentées, mais nous verrons qu'elles ne sont pas entièrement adaptées à la problématique de l'anonymisation. Enfin, les recherches actuelles concernent à présent l'hybridation des méthodes symboliques et à base d'apprentissage statistique. Ce type de démarche donne, à l'heure actuelle, les résultats les plus prometteurs en la matière.

❧

Dans un second temps, nous introduisons la question de l'évaluation des résultats, en nous plaçant dans le contexte d'outils d'anonymisation automatique. Cette mise en perspective implique d'évaluer deux dimensions des systèmes informatiques, la catégorisation de l'information traitée d'une part, et la délimitation de la portion porteuse d'information d'autre part. Nous passerons en revue les différentes mesures qui existent pour évaluer les résultats de l'anonymisation. Nous présenterons également les mesures utilisées pour évaluer les annotations humaines, et notamment les accords inter-annotateurs. Enfin, puisque les outils ne sont généralement appliqués que sur un nombre restreint de documents — à hauteur du nombre de documents constituant le corpus de test —, nous verrons par quel moyen il est possible de simuler les résultats qu'obtiendraient ces outils s'ils étaient appliqués sur un plus grand nombre de documents.

Chapitre 2

Méthodologies

Le bibliothécaire avait vu pas mal de spectacles insolites dans sa vie, mais celui-là arrivait indubitablement en cinquante-septième position (il avait l'esprit méthodique).

Les zinzins d'Olive Oued
TERRY PRATCHETT

Sommaire

2.1 Introduction

Dans ce chapitre, nous passons en revue les différentes méthodes existantes permettant de traiter de la problématique de l'anonymisation automatique de documents du domaine médical.

À l'image des approches appliquées en fouille de textes, les approches utilisées en anonymisation automatique reposent sur deux grandes familles de méthodes : les méthodes à base de règles *(généralement implémentées sous la forme d'expressions régulières)* et de listes *(listes d'entités, dictionnaires, etc.)*, dites « méthodes symboliques », et les méthodes à base d'apprentissage statistique reposant sur la construction d'un modèle à partir d'un corpus annoté. La combinaison de ces deux familles de méthodes donne lieu aux méthodes hybrides.

Cependant, quelle que soit la méthode choisie, il importe de se rappeler que le système développé reste fortement lié aux caractéristiques du corpus pour lequel il a été réalisé, à plus forte raison s'il s'agit d'un système à base d'apprentissage. Comme le soulignent [Ferrández et al., 2012], en matière d'apprentissage, le modèle construit sur un corpus est difficilement applicable sur un autre corpus.

Nous inscrivons notre présentation des approches existantes dans le domaine médical dans un double objectif : en premier lieu, l'anonymisation des données personnelles contenues dans les documents cliniques de telle sorte que la réidentification du patient soit la plus difficile possible, et deuxièmement, le maintien d'une qualité finale du corpus anonymisé suffisante pour permettre l'application de traitements ultérieurs.

Une partie importante de ce chapitre repose sur l'édition 2006 du défi i2b2, consacré à la problématique de l'anonymisation automatique de comptes rendus cliniques rédigés en anglais. La confrontation de plusieurs systèmes sur le même jeu de données nous autorise une comparaison directe des systèmes utilisés par les différents participants. Plus récemment, certaines équipes de recherche ont comparé l'application de leur propre système d'anonymisation avec les outils existants, appliqués sur les mêmes corpus. Il importe par ailleurs de souligner qu'en anglais, le corpus i2b2 2006, l'un des rares corpus annotés disponibles, est très fréquemment utilisé à cet effet.

2.2 Les méthodes à base de règles

2.2.1 Présentation

Les méthodes à base de règles sont coûteuses à implémenter en termes de temps, du fait de la mobilisation des connaissances d'experts qu'elles nécessitent, et se révèlent également coûteuses lors des étapes de maintenance et d'extension des règles à de nouveaux domaines. Elles offrent néanmoins des résultats de qualité, pour peu que les experts ayant produit ces règles maîtrisent les moyens de les implémenter d'une part *(expressions régulières, algorithmes informatiques, etc.)*, et l'ensemble des formes possibles de présentation des informations d'autre part. De plus, contrairement aux approches par apprentissage statistique, elles ne nécessitent pas obligatoirement de corpus annoté ; l'existence d'un corpus annoté peut néanmoins se révéler utile pour tenter d'inférer automatiquement des règles.

Les inconvénients majeurs de ce type d'approche concernent la difficulté d'adaptation de ces règles à de nouveaux types de documents ou à des informations non

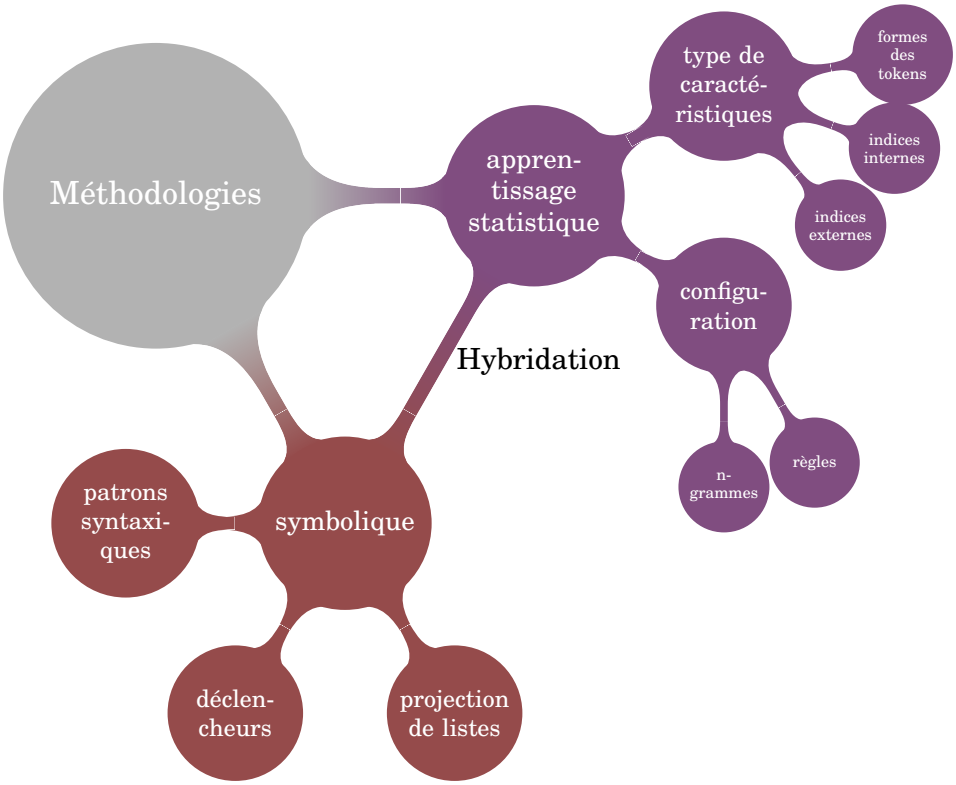


FIGURE 2.1 – Méthodologies employées pour l'anonymisation automatique

couvertes. Les règles souffrent en effet d'un manque de robustesse qui ne permet, en règle générale, pas de les appliquer à un autre domaine. L'adaptation à un nouveau domaine ou à un nouveau type de document se révélera alors coûteuse (du point de vue de la mobilisation d'expert qu'elle implique). D'autre part, pour que les règles soient efficaces, il convient que les experts qui les définissent soient conscients de toutes les formes possibles de présentation des informations traitées.

L'utilisation des méthodes à base de règles est donc optimale sur des documents présentant les caractéristiques suivantes : des documents de même origine *(un seul hôpital, un seul service au sein de cet hôpital)*, issus d'un traitement de texte *(des documents correctement rédigés et bien formatés)*, et dans une moindre mesure, des documents de même type *(lettres, comptes rendus, résultats d'examen, etc.)*. On appliquera ainsi ce type de méthodes sur un corpus de documents homogènes.

A contrario, il ne sera guère possible d'utiliser les méthodes à base de règles sur des documents issus d'une numérisation, en raison des erreurs de reconnaissance de caractères qu'implique ce processus *(ajout ou suppression d'espaces, ajout de caractères exotiques, transformation de caractères, etc.)*, du fait de l'absence de robustesse des règles.

Le fonctionnement global d'un système à base de règles repose sur deux types d'éléments particuliers : des ressources externes et des règles (voir figure 2.2).

Ressources externes

Sous le vocable « ressources externes », on regroupe généralement deux catégories de ressources.

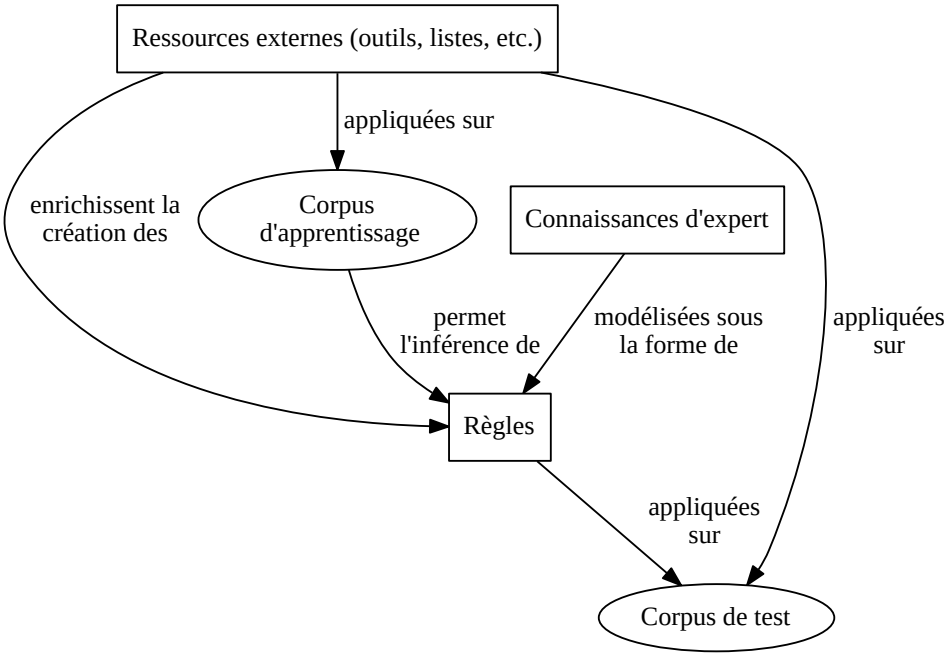


FIGURE 2.2 – Architecture générale d'un système symbolique

– Des ressources directement exploitables telles que des listes de termes *(listes de noms, de prénoms, de villes, etc.)*, des listes de déclencheurs [1] *(déclencheurs de personnes : « M., Mme, Melle, Dr, Pr », etc.)* et des dictionnaires *(langue générale ou langue de spécialité)* ;
– Des outils externes qui permettent d'apporter une information supplémentaire *(étiqueteur-lemmatiseur, analyseur syntaxique, etc.)* qui sera ensuite exploitée sur le corpus.

Ces ressources externes peuvent être utilisées de deux manières. En premier lieu, il est possible de projeter directement le contenu de ces ressources sur les corpus pour identifier, à l'identique, des éléments du texte présents dans l'une de ces listes (exemple 4).

(4) Projection de la liste des villes sur le corpus : *75006* [ville *Paris*].

En second lieu, ces ressources peuvent servir à un humain pour inférer des règles depuis le corpus *(si l'on remarque des régularités liées à l'utilisation de ces ressources, exemple 5)*. Elles peuvent également être utilisées en combinaison avec des règles de manière à enrichir ces dernières (exemple 6).

(5) Observation récurrente permettant à un humain d'inférer une règle, les intitulés « Opérateur », « Assistants » et « Anesthésiste » présents dans l'ensemble des comptes rendus opératoires d'un corpus introduisent le nom de médecins :
Opérateur : Pr [nom *Martin*]
Assistants : DR [nom *Bernard*] *+ DR* [nom *Dupont*]

1. On appelle « déclencheur » un élément du texte servant d'indice pour détecter une entité nommée.

Anesthésiste : DR [nom *Belleville*] .

(6) Un terme absent du dictionnaire de langue, commençant par une capitale, précédé d'un déclencheur de personne et/ou suivi d'un élément présent dans la liste des prénoms a de fortes chances d'être un nom : *J'ai effectué la tomoscintigraphie myocardique à l'effort de Monsieur* [nom *Stipe*] *Michael âgé de 71 ans.*

Expressions régulières

Les connaissances sur la composition d'une entité ou sur son contexte habituel — qu'elles soient d'ordre général ou qu'elles relèvent d'un domaine scientifique particulier — sont modélisées et implémentées sous la forme d'expressions régulières. Ce type d'approche permet particulièrement bien de traiter des cas numériques (exemples 7 et 8) car il est facile et rapide d'en modéliser toutes les formes possibles.

(7) En France, un code postal se compose de cinq chiffres : [code-postal *75006*] *Paris.*

(8) Une mesure se compose d'une valeur (un entier ou une décimale) suivie d'une unité de mesure : *Détensiel* [mesure *10 mg/j*] .

Dans une moindre mesure, les règles permettent de traiter les chaînes de caractères dans des contextes bien particuliers (exemple 9).

(9) Une chaîne de caractères précédée de la mention « *hospitalisation (dernière | récente |) (à | au)* » et suivie d'une ponctuation ou d'un mot grammatical (article, préposition, etc.) a de fortes chances d'être un nom d'hôpital : *hospitalisation au* [hôpital *centre hospitalier de Saint-Malo*] *le 5 juin 1993.*

Sur les chaînes de caractères, si elles ne sont pas utilisées en combinaison avec le contenu de listes ou de dictionnaires, les règles sont plus complexes à produire et ne permettent pas de couvrir tous les cas de figures possible.

2.2.2 Approches suivies

[Meystre et al., 2010] relèvent deux étapes principales dans les approches symboliques :

- Une première étape consiste à appliquer des patrons syntaxiques qui sont implémentés sous la forme d'expressions régulières, et des déclencheurs. Ces patrons servent essentiellement pour les informations numériques. Une éventuelle définition de priorité entre règles peut être réalisée, notamment pour gérer les cas d'enchâssement de portions avec une priorité accordée aux portions les plus vastes au détriment des règles locales ;
- La seconde étape revient à projeter des listes *(par exemple issues de recensements)* pour les noms, les prénoms et les lieux.

En complément de ces deux étapes partagées par tous les systèmes symboliques s'ajoutent des étapes supplémentaires de pré-traitements ou de post-traitements.

Approches basiques

Les approches les plus simples consistent à projeter des listes de termes sur les corpus, et à compléter cette projection par des règles. Tous les outils fondés sur des approches symboliques reposent sur ces deux étapes.

[Fielstein et al., 2004] ont produit un script Perl qui repose sur des expressions régulières. Les patrons syntaxiques ont été créés à partir de ressources publiques. Ils ont fait l'objet d'une adaptation aux besoins du corpus sur lequel cet outil a été appliqué, en l'occurrence, un corpus de rapports d'examens de pensions de l'institut Veterans Affairs aux États-Unis *(VA Pension examination reports)*. Sur ce corpus spécifique, les auteurs rapportent un rappel assez bon (0,81) ainsi qu'une spécificité élevée *(taux de vrais négatifs)* (0,99), ce qui n'est pas particulièrement informatif.

Après avoir balisé les informations relevant des plans personnels et pathologiques, [Beckwith et al., 2006] ont appliqué des expressions régulières puis recherché à l'identique les éléments présents dans les données issues de recensement. Leur outil, baptisé « HMS Scrubber » repose sur les étapes suivantes : (i) une conversion du document au format XML est d'abord réalisée, comprenant notamment des balises d'entête — pour les informations personnelles *(nom, prénom, adresse, numéro de Sécurité Sociale)* et pathologiques *(département médical)* sur le patient — et de texte, (ii) l'outil effectue une application d'expressions régulières (50 patrons avec déclencheurs), et (iii) recherche à l'identique dans une base de données contenant des noms, prénoms, lieux provenant de recensements. L'approche suivie leur a permis d'obtenir un très bon rappel (0,98) au détriment d'une précision assez faible (0,43), autrement dit, presque toutes les informations devant être anonymisées l'ont bien été, en revanche, l'outil a procédé à une sur-anonymisation importante.

Exploitation des entêtes de document

Une approche sensiblement identique est suivie dans le cadre du système commercial « De-Id » [Gupta et al., 2004] développé à UPMC. [2] Des expressions régulières sont utilisées pour les informations numériques, complétées par des règles, des dictionnaires et des listes issues des recensements. Une amélioration notable de ce système repose sur la réutilisation des informations issues de l'entête du document *(informations administratives sur le patient, nom de médecin, département clinique)*, approche également suivie par [Friedlin et McDonald, 2008]. Ces informations sont alors projetées sur le reste du document pour faciliter la détection des informations relatives au patient. Enfin, le métathesaurus de l'UMLS est utilisé pour repérer les termes médicaux qui sont alors maintenus en clair dans le document car étant considérés comme porteurs d'informations. Les informations anonymisées sont remplacées par des étiquettes typantes avec une propagation des identifiants dans le document : le même nom sera toujours remplacé par la même étiquette tout au long du document, les dates sont remplacées par d'autres qui maintiennent l'écart de temps.

Utilisation d'informations sémantiques

Une autre approche, plus basique, est celle suivie par [Morrison et al., 2009a]. Cette approche consiste à d'abord distinguer les termes médicaux des termes non médicaux grâce à une vérification de chaque terme dans des listes, notamment le contenu de l'UMLS [3] [Lindberg et al., 1993].

Selon un principe équivalent, [Berman, 2003] a préalablement effectué une recherche des termes du document dans l'UMLS afin de conserver en clair ces éléments, considérant qu'ils sont porteurs d'information. Son outil — intitulé « concept-match

2. University of Pittsburgh Medical Center, Pittsburgh, PA.
3. http://www.nlm.nih.gov/research/umls/

scrubber » – est un script Perl qui repose sur les étapes suivantes : (i) l'outil effectue un parcours du document pour effectuer une segmentation $a)$ en phrases, $b)$ en mots, et pour identifier les mots à ne pas traiter *(stop-words)* ; (ii) une recherche à l'identique des mots pleins dans le métathesaurus UMLS est alors effectuée, en effectuant prioritairement une recherche des portions les plus longues, ce qui conduit à remplacer les termes identifiés par l'identifiant CUI *(Concept Unique Identifier)* et le terme vedette correspondant présent dans l'UMLS. Enfin, (iii) le remplacement des termes qui n'ont pas été trouvés est réalisé au moyen d'une étiquette bloquante. Les auteurs rapportent que ce type d'approche conduit à une faible précision, autrement dit, un taux de suranonymisation élevé.

Utilisation de listes

L'utilisation de dictionnaires de mots communs et médicaux permet de gérer l'absence inhérente d'exhaustivité des listes. C'est notamment l'approche suivie par [Thomas et al., 2002] dont le système traite exclusivement de l'anonymisation des noms propres. Un terme sera anonymisé : (i) s'il est directement présent dans une liste de noms propres, ou (ii) si le voisinage de ce terme est un déclencheur *(M., Mme, Melle)*, ou encore (iii) si le terme n'est présent dans aucune liste *(ni dans la liste des noms propres, ni dans le dictionnaire de termes communs)* mais que son voisinage contient un nom propre. Les listes de noms utilisés proviennent de trois sources distinctes : les noms propres du correcteur orthographique *Ispell*,[4] les noms de médecins et de patients de la base de données de l'institut Regenstrief, et les noms issus de l'index des décès de la Sécurité Sociale américaine. Les auteurs rapportent un rappel élevé (0,987). La prise en compte du contexte en complément des listes est également l'approche qui a été suivie par [Sibanda et Uzuner, 2006].

Pour développer leur outil « De-ID »[5] au sein du MIT,[6] [Neamatullah, 2006, Neamatullah et al., 2008] ont utilisé le contenu de la base de données MIMIC II[7] *(données cliniques et unités de soins intensifs en cardiologie aux USA issues du corpus Physionet)* pour identifier les noms des patients et des médecins. Ils ont également produit deux listes pour distinguer : (i) les noms et lieux ambigüs devant être anonymisés *(présents à la fois dans l'UMLS et dans le dictionnaire de noms communs)*, (ii) de ceux (termes communs et médicaux) ne devant pas être anonymisés. Il importe de noter que cet outil, développé en Perl, est l'un des rares outils à être librement téléchargeable[8] et modifiable. Pour cette raison de disponibilité, certaines équipes ont repris et modifié cet outil pour l'adapter aux spécificités de leur corpus *(un corpus de comptes rendus cliniques sur les soins primaires)* [Tu et al., 2010].

Une autre approche à base de listes repose sur l'hypothèse qu'une relation forte existe entre certaines classes de mots et certaines classes de concepts. Partant de cette hypothèse, [Taira et al., 2002] ont mis au point un système d'identification des noms propres fondé sur les relations grammaticales *(appelées « relations logiques »)* présentes dans chaque phrase et une liste de 64 000 noms et prénoms. Une relation

4. http://fmg-www.cs.ucla.edu/geoff/ispell.html
5. Attention : il existe deux systèmes nommés « De-ID », le premier est commercial, réalisé par [Gupta et al., 2004] à l'Université de Pittsburgh, tandis que le second, réalisé par [Neamatullah, 2006, Neamatullah et al., 2008] au MIT, est librement réutilisable. Tous deux reposent sur des approches symboliques.
6. Massachusetts Institute of Technology, Cambridge, MA.
7. http://mimic.physionet.org/
8. http://www.physionet.org/physiotools/deid/

logique consiste en un prédicat, qui indique le type de relation *(Patient-healthStatus, Patient-condition, Patient-procedure)*, et une liste composée d'un ou plusieurs arguments. Dans la phrase *Johnny underwent a pyeloplasty for uretropelvic junction stenosis*, le prénom *Johnny* est la tête de la relation, *underwent* la relation et *pyeloplasty* la valeur ; le prédicat de la relation logique est de type *Patient-procedure*. Appliqué sur les noms propres, les auteurs rapportent une précision élevée (0,99) et un très bon rappel (0,94). Notons que leur système se limite uniquement à cette catégorie.

Précédence et pondération

Alors que la majorité des approches repose sur l'utilisation de listes d'entités nommées combinées à des patrons syntaxiques, [Sweeney, 1996] a mis au point une méthode tenant compte des notions de « précédence » *(la règle de détection des lieux prime sur celle des villes)* et de « pondération des règles » *(la règle qui possède la plus forte précédence et dont le score est le plus élevé prévaut sur les autres)*. Plus une entité est constituée de composants *(d'autres entités de plus bas niveau)*, plus son score de précédence sera élevé.[9] Leur outil, « Scrub », effectue un remplacement des informations anonymisées par des informations vraisemblables. L'évaluation des résultats produits aboutit à un rappel élevé (0,99) mais ne fournit aucune indication sur la précision et du taux de sur-anonymisation. Appliquer des règles de précédence permet de régler les problèmes d'anonymisation partielle et conduit, si elle est envisagée correctement dès la création de l'anonymiseur, à une amélioration de la précision. Notons que ces notions de précédence et de pondération jouent sur la procédure de détection finale. En cela, elles sont à mettre en parallèle des méthodes par apprentissage (section 2.3).

Distance d'édition

Dans leur système « MeDS », en plus des habituelles listes et expressions régulières, [Friedlin et McDonald, 2008] ont également implémenté une distance d'édition pour repérer les fautes de frappe sur les noms et prénoms. Cette implémentation leur permet d'obtenir un excellent rappel (0,995 sur les identifiants HIPAA et 0,969 sur les identifiants hors HIPAA) avec une sur-anonymisation limitée, évaluée à 8 %.

Étiquetage en parties du discours

Sur le français, [Ruch et al., 2000] ont mis au point un algorithme — intitulé « MedTag » — qui repose : (i) sur un étiquetage en parties du discours des différents éléments du texte, combiné (ii) à une identification de la classe sémantique de ces éléments, sur la base d'un lexique sémantique spécialisé en médecine. Sur la base de ce double étiquetage, (iii) une décision est alors prise en matière d'anonymisation. Sur l'exemple 10 issu d'un article rédigé en anglais, un réseau de transition récursif prend alors la décision d'anonymiser les noms *Smith* et *John*. Les auteurs rapportent le fait qu'aucune sur-anonymisation n'a été produite par leur outil.

(10) [nom *Doctors*] [nom *Smith*] *and* [nom *John*] *saw the* [person *patient*] .

9. L'entité *address block* regroupe cinq entités (*street, city, state, zip* et *country*) alors que l'entité *location* n'en regroupe que trois (*city, state* et *country*), le score de précédence de l'entité *address block* (5) est donc supérieur à celui de l'entité *location* (3), lui-même supérieur à celui des autres entités listées comme composants (1).

Conclusion

Des différentes méthodologies suivies par ces différents outils, nous estimons que les deux approches suivies par l'outil De-ID *(utiliser des listes et des expressions régulières ; distinguer les informations à anonymiser des informations ne devant pas l'être)* paraissent intéressantes dans le traitement de l'anonymisation automatique. Nous envisageons d'exploiter autant que possible ces deux approches dans un système d'anonymisation automatique en combinaison avec les méthodes traditionnelles des approches symboliques *(définition de règles, projection de lexiques et utilisation de déclencheurs)*.

La réutilisation d'informations explicites sur le patient, telles que celles que l'on peut trouver dans l'entête des documents cliniques ou en interrogeant le système d'information patient (SIP) de l'hôpital, permet de s'assurer d'un traitement correct [10] des informations nominatives et numériques relatives au patient, s'agissant en l'occurrence des informations les plus sensibles dans un document clinique en matière d'anonymisation.

2.2.3 Pallier l'absence de listes

L'un des moyens de pallier l'absence de listes, ou l'absence d'exhaustivité des listes existantes, consiste à tenter de formaliser les règles qui régissent la formation de certaines catégories de mots en se fondant notamment sur certains critères « internes » (voir p. 86).

Si cela s'avère difficile et incomplet pour les mots de la langue courante *(lister certains suffixes pour formaliser la classe des adjectifs ou des substantifs)*, il existe une catégorie d'informations relative aux corpus médicaux pour laquelle il est envisageable de définir des règles de formation des éléments qui la composent : les noms de médicaments.

La composition d'un médicament comprend, en règle générale, la substance active et un ensemble d'autres ingrédients. La substance active *(également appelée « molécule de base »)* procure au médicament son efficacité dans le soin visé et détermine la classe à laquelle appartient le médicament. Les autres ingrédients ne servent que dans bien des cas à modifier le goût et la couleur du médicament, de manière à le rendre plus supportable pour les malades ; ces autres ingrédients interviennent également dans la formule globale qui permet au médicament de se distinguer de ceux produits par la concurrence ou des médicaments dits « génériques ».

Les noms de substances

Depuis 1953, l'organisation mondiale de la santé (OMS) élabore une dénomination commune internationale (DCI). L'objectif de cette DCI consiste à lister les différentes substances actives de manière universelle, codifiée, et facile à mémoriser. Dans cette perspective, des affixes appelés « segments-clés » ont été définis pour rendre compte de cette classification. Les divers organismes ou commissions de nomenclature nationaux ont été associés pour uniformiser les appellations existantes et pour définir les nouvelles DCI.

10. Pour peu que les informations aient été correctement entrées dans le SIP *(sans faute de frappe : Gabriel vs. Gabrielle ; 2006 vs. 2OO6)* et qu'aucun problème d'encodage des caractères ne se pose.

Les segments-clés. La majorité des segment-clés est créée à partir d'un élément du nom de la substance qui constitue la famille pharmaco-thérapeutique (voir tableau 2.1). Ces segment-clés peuvent apparaître en position initiale (préfixe), médiane (infixe), ou finale (suffixe).

Position	Segment-clé	Latin	Famille pharmaco-thérapeutique	Exemples de substances
Initiale	io-	*io-*	Produits de contraste iodés	*Iopamidol, Iopromide*
	vin-	*vin-*	Alcaloïdes de la pervenche Vinca.	*Vincamine, Vinblastine*
Médiane	-prost-	*-prost-*	Prostaglandines et leurs dérivés.	*Alprostadil, Latanoprost*
	-rétin-	*-retin-*	Dérivés du rétinol (Vitamine A).	*Etrétinate, Isotrétinoïne*
Finale	-ac	*-acum*	Anti-inflammatoires non stéroïdiens dérivés de l'ibufénac.	*Diclofénac, Bufexamac*
	-mycine	*-mycinum*	Antibiotiques produits par diverses souches de Streptomyces.	*Erythromycine, Spectinomycine*

TABLE 2.1 – Segment-clés créés à partir d'affixes de noms de substances.

Certains segments-clés sont formés par la contraction de deux mots renvoyant à la famille pharmaco-thérapeutique à laquelle renvoie le segment-clé ainsi créé (voir tableau 2.2).

Par ailleurs, certains segment-clés génériques se trouvent sous-spécifiés et donnent lieu à de nouveaux segment-clés intégrant le segment-clé générique. Nous donnons dans le tableau 2.3 l'exemple des segment-clés utilisés dans le cas des antiviraux.

Les classifieurs suffixoïdes. [Corbin et Paul, 2000] appellent « classifieurs suffixoïdes » les segments-clés présents sous forme de suffixes utilisés pour la classification d'éléments : « *Il s'agit de finales suffixiformes qui ont, sémantiquement, un rôle de classifieurs.* » Les auteurs précisent que ces classifieurs suffixoïdes sont spécifiques à certaines terminologies *(chimie, pharmacologie, etc.)* et n'ont pas d'équivalents dans le lexique commun. Sur la base de ces classifieurs suffixoïdes et par extrapolation, il est également possible de déterminer, d'après la DCI, des « classifieurs préfixoïdes » et des « classifieurs infixoïdes » :

- Le préfixe *vin-* désigne les alcaloïdes de la pervenche Vinca. Il est utilisé dans le nom de substance *vincamine*, substance que l'on retrouve dans le médicament *pervincamine*.
- L'infixe *-retin-* provient du terme *rétinol* lequel renvoie à la vitamine A. Cet infixe est utilisé dans les noms des dérivés de la vitamine A, tels que l'*isotrétinoïne* (présente dans l'*isotrex* ou le *roaccutane*) ou l'*étrétinate* (présente dans le *tigason*).
- Le suffixe *-ac* provient du nom de la substance *ibufénac*. Il désigne tous les anti-inflammatoires non stéroïdiens dérivés de l'*ibufénac*. On trouve ce segment-clé

Segment-clé	Famille pharmaco-thérapeutique	Exemples de substances
-bactam	Antibiotiques inhibiteurs des **b**êta-l**actam**ases.	*Sulbactam*
-bendazole	Anthelminthiques à noyau **ben**zimi**dazole**.	*Albendazole, Mébendazole*
-carbef	Antibiotiques dérivés des **carb**ac**éph**ems du groupe des bêta-lactamines.	*Loracarbef*
-nidazole	Anti-infectieux et anti-parasitaires à **n**oyau im**idazole**.	*Métronidazole, Ornidazole, Secnidazole, Tinidazole*
-onidine	Antihypertenseurs, guanidines à **n**oyau imidaz**olidine**.	*Apraclonidine, Clonidine*
-profène	Anti-inflammatoires non stéroïdiens dérivés de l'acide **phén**yl**pro**pionique.	*Alminoprofène, Flurbiprofène, Ibuprofène, Kétoprofène*
-stérone	**Stér**oïdes cét**on**iques divers (androgènes, progestatifs, hormonaux).	*Prastérone, Dydrogestérone, Progestérone, Aldostérone*

TABLE 2.2 – Segment-clés créés par la contraction de mots.

dans le nom de la molécule *bufexamac*. Cette molécule est présente dans le médicament *parfenac*, lequel intègre également dans son nom le segment-clé.

Les noms de médicaments

Contrairement aux noms de substances qu'il est aisé d'identifier dans des textes sur la base de ces précédents classifieurs, les noms de médicaments se composent : (i) du segment-clé correspondant à la classe à laquelle appartient le médicament ; et/ou (ii) de la cible visée par le médicament, c'est-à-dire soit l'organe concerné, soit le mode d'administration. Un médicament étant un produit destiné à être commercialisé, le nom peut également être élaboré de la même manière qu'un nom de marque, sans que la référence au segment-clé ou à la cible visée ne soit apparente.

- Un exemple de nom fondé sur l'intégration du segment-clé concerne le médicament *néo-cort*. Le préfixe *cort-* renvoie aux corticostéroïdes autres que les dérivés de la prednisolone. Ce segment-clé apparaît sous forme d'infixe dans le nom de la substance active *hydrocortisone*, substance présente dans le médicament *néo-cort*. Dans cet exemple, le nom du médicament reprend le segment-clé utilisé dans la substance active ;
- À l'inverse, rien dans le nom du médicament *oro-pivalone* ne fait référence à un quelconque segment-clé. Il s'agit d'une création dont le préfixe renvoie au mode d'administration *(oro = par voie orale)*. Le nom du principe actif du médicament *(tixocortol)* contient le segment-clé *cort-*.

Segment-clé	Famille pharmaco-thérapeutique	Exemples de substances
-gest-	Stéroïdes progestatifs (non cétoniques)	*Nomégestrol, Promégestone*
-stérone	Stéroïdes cétoniques divers	*Prastérone, Aldostérone*
-gestérone	Stéroïdes cétoniques progestatifs	*Dydrogestérone, Progestérone*
-vir	Antiviraux divers	*Aciclovir, Ganciclovir*
-amivir	Antiviraux inhibiteurs de la neuraminidase	*Oseltamivir, Zanamivir*
-cavir	Antiviraux nucléosides carbocycliques	*Abacavir*
-navir	Antiviraux inhibiteurs de la protéase du VIH	*Indinavir, Saquinavir*

TABLE 2.3 – Segment-clés sous-spécifiés.

2.2.4 Les méthodes à base de repérage d'entités nommées

Présentation

Depuis leur lancement en 1987, les campagnes MUC *(Message Understanding Conference)* — financées par le DARPA [11] — se focalisent sur des tâches d'extraction d'information depuis des textes non structurés, typiquement des articles de journaux ou des dépêches d'agences de presse, relevant de domaines thématiques tels que la défense ou les activités commerciales. C'est lors de la campagne d'évaluation MUC-6 qu'a été introduite pour la première fois la notion d'« entités nommées » [Grishman et Sundheim, 1996]. À l'occasion de cette édition, il a été considéré comme essentiel de repérer dans les textes journalistiques du *Wall Street Journal* les unités d'information présentes sous la forme de noms propres. Depuis cette campagne, [12] les entités nommées sont donc considérées comme des noms propres relevant de trois catégories principales : les noms de *personnes*, les noms de *lieux*, et les noms d'*organisations* (groupe des ENAMEX, *Entity Name Expressions*) [Sekine et Ranchhod, 2009]. Des expressions numériques ont également été intégrées sous cette notion pour couvrir les dates et les heures (groupe des TIMEX, *Time Expressions*), les montants et les pourcentages (groupe des NUMEX, *Number Expressions*). Le repérage des entités nommées (REN) est la tâche qui consiste à identifier dans des textes les références à ces entités. Autrement dit, à identifier les entités puis à catégoriser ces entités. Le REN est une sous-tâche particulièrement importante de l'extraction d'information.

Cette représentation de base des entités nommées a par la suite donné lieu à de nombreux travaux, soit pour étendre la couverture des entités, soit pour apporter des précisions sur les catégories de base. On peut ainsi dégager deux types d'améliorations effectuées ces dernières années :

– L'augmentation du nombre de catégories principales en ajoutant de nouvelles réalités : [Sekine, 2004] a notamment défini une hiérarchie complète d'entités nommées, composée d'environ 200 types, ou bien l'ajout de catégories précises pour couvrir une thématique particulière *(les gènes et les protéines, une période*

11. http://www.darpa.mil/, Defense Advanced Research Project Agency, Arlington, VA.
12. MUC-6 : http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

particulière de l'Histoire, etc.) ou un type de documents précis *(un corpus de messages électroniques)* ;
– La subdivision des catégories existantes en intégrant des sous-classes : les noms d'hommes *politiques* sous la classe des « personnes », les noms de *villes* sous la classe des « lieux », etc. [Fleischman et Hovy, 2002]. L'ajout de ces sous-classes permet une représentation sémantique plus fine du contenu des documents ainsi qu'un accès plus précis au contenu des documents.

Le résultat d'un REN peut donner lieu à de nombreuses applications en traitement automatique des langues, telles que celles renseignées par [Ehrmann, 2008] : (i) un pré-traitement pour de futures étapes *(analyse syntaxique, résolution de coréférences, etc.)*, (ii) fournir une aide contextuelle à la désambiguïsation sémantique, pour permettre la distinction des différents sens d'un verbe en fonction de la présence d'une entité nommée dans le contexte *(« quitter Paris » vs. « quitter quelqu'un/quelque chose »)*, et (iii) fournir une aide pour la traduction automatique, par exemple pour distinguer le nom de famille « London » dans « Jack London » du nom de la capitale britannique. Sur ce dernier point, [Ehrmann, 2008] rapporte la traduction *« Jack Londres était un auteur américain. »* fournie par le logiciel Systran pour la phrase d'origine *« Jack London was an american writer. »*.

En matière de fouille de textes, le repérage des entités nommées sert traditionnellement : (i) à répondre à des questions basiques *(Qui ? Quoi ? Où ? Quand ? Comment ?)*, ou (ii) à peupler une base de données *(base de faits d'actualités depuis des dépêches d'agence de presse)*. Le REN peut également constituer une application dont l'anonymisation serait la conséquence.

Les entités nommées pour l'anonymisation automatique

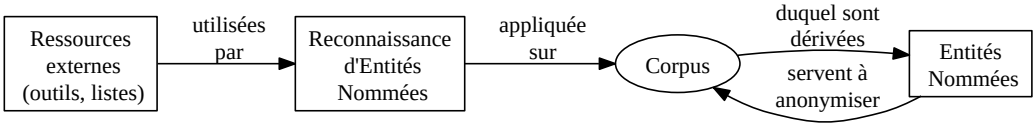


FIGURE 2.3 – Utilisation d'un système de reconnaissance d'entités nommées pour anonymiser un corpus

Il est possible d'utiliser et d'adapter des outils de Repérage d'Entités Nommées pour effectuer une anonymisation (figure 2.3). [Aberdeen et al., 2010] rapportent que quatre heures de travail seulement se sont révélées nécessaires aux organisateurs de l'édition 2006 du challenge i2b2 pour adapter le système Carafe [13] – un outil à base de CRF initialement développé pour effectuer du repérage d'entités nommées [Wellner, 2009] — aux besoins du challenge.

Dans le domaine juridique, des travaux ont porté sur l'anonymisation de décisions de justice canadiennes dans le but d'offrir au grand public un accès aux décisions de jurisprudence, tout en garantissant le respect de la vie privée. [Plamondon et al., 2004] ont ainsi développé un prototype à base de REN fonctionnant avec la plate-forme GATE pour anonymiser les noms de personnes, de lieux

13. http://mist-deid.sourceforge.net/docs_1_2/html/carafe_engine.html

et d'organisations. Une évaluation réalisée sur les 63 formes (total des formes d'entités à anonymiser sur les trois catégories), et non sur les 546 occurrences des entités à anonymiser, renvoie un rappel de 0,73 et une précision de 0,88 (soit une F-mesure de 0,80). Ce module d'anonymisation a également fait l'objet d'un développement sous la forme d'une macro intitulée « NOME » pour le traitement de texte Word [Pelletier et al., 2004].

Plus récemment et pour le français, [Gicquel et al., 2012] ont adapté un outil de REN utilisé lors de la campagne d'évaluation ESTER [14] pour effectuer une anonymisation en milieu hospitalier pour la langue française. Le corpus utilisé se compose de 1 000 documents écrits *(comptes rendus hospitaliers, comptes rendus opératoires, comptes rendus d'imagerie, lettres de consultation et lettres de suivi)* provenant de quatre Centres Hospitaliers Universitaires français, et pour chacun de services différents *(chirurgie digestive, chirurgie orthopédique, neurochirurgie, et réanimation)*. L'anonymisation a été effectuée au sein de chaque CHU au moyen de l'outil d'anonymisation ainsi développé.

Cet outil repose sur deux étapes : (i) une étape de repérage d'entités nommées automatique, sur la base d'un outil existant ayant été adapté au domaine médical, et (ii) une étape de visualisation via une interface graphique permettant l'annotation des éléments non repérés au terme de la première étape.

Les auteurs mentionnent quatre niveaux d'adaptation de cet outil de REN : (i) une adaptation lexicale *(liste d'abréviations du domaine, etc.)*, (ii) une adaptation de la désambiguïsation des parties du discours *(regrouper sous une seule entité linguistique un nom de maladie composé de plusieurs tokens et gérer les interprétations multiples de termes)*, (iii) une normalisation du format des documents traités *(notamment les absences de point de fin de ligne)*, et (iv) un traitement spécifique appliqué aux dates, de manière à conserver l'écart temporel entre la date d'hospitalisation *(définie comme T0)* et les autres dates présentes dans le document *(alors définies comme T+1J pour le lendemain de l'hospitalisation)*.

La performance globale de l'outil est relativement bonne (Rappel=0,797, Précision=0,852 et F-mesure=0,824). Dans le détail, les auteurs rapportent des valeurs élevées pour l'anonymisation des données numériques (dates R=0,955, F=0,976 et téléphones R=0,810, F=0,895) ou des données aisément formalisables (adresses électroniques R=0,952, F=0,976). Les valeurs sont moins élevées sur les noms et prénoms en générale (R=0,827, F=0,829) et chutent si une désambiguïsation est effectuée entre patient (R=0,825, F=0,703) et professionnel de la santé (R=0,706, F=0,793). Le traitement de la catégorie des lieux n'a pas permis l'obtention de résultats élevés (R=0,449, F=0,497).

Cependant, [Benton et al., 2011] considèrent que l'extension des méthodes de REN pour effectuer une anonymisation ne suffit pas, en raison des spécificités inhérentes aux documents cliniques *(erreurs typographiques, plus grande variété de noms, utilisation de termes ou d'abréviations particulières, etc.)* qui induisent des difficultés supplémentaires.

14. http://www.afcp-parole.org/camp_eval_systemes_transcription/, ESTER, Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques.

2.3 Les méthodes à base d'apprentissage statistique

2.3.1 Présentation

L'anonymisation automatique au moyen d'approches à base d'apprentissage statistique est généralement perçue, soit (i) comme une tâche de reconnaissance d'entités nommées (REN) avec des outils effectuant une identification de l'entité *(repérage de l'entité et définition des bornes de l'entité)* et un typage de l'entité *(affectation du label de catégorie à laquelle appartient l'entité)*, soit (ii) comme une tâche de classification dans le sens où chaque token sera classé dans une ou plusieurs classes parmi les catégories d'anonymisation définies. Cependant, dans le cas où l'approche est perçue comme relevant d'un repérage d'entités nommées, [Uzuner et al., 2007, p. 551] relèvent que l'ambiguïté de certains termes et l'absence de certaines informations dans des listes réduit la contribution des dictionnaires et des listes. Les auteurs indiquent que l'étude du contexte et des particularités de la langue médicale permettent de pallier ces lacunes. [15]

Définition

Nous reprenons la définition d'« apprentissage artificiel » telle que formulée par [Wisniewski, 2007] : « *l'apprentissage artificiel a pour objectif la mise au point de programmes capables d'apprendre à partir de leur expérience, c'est-à-dire de changer leur structure interne ou la valeur de leurs paramètres en fonction de leur expérience de manière à améliorer leurs performances futures.* »

L'utilité des méthodes à base d'apprentissage artificiel apparaît dans deux cas précis : (i) lorsque les programmes qui doivent être développés sont trop complexes à coder, en particulier si de nombreux paramètres doivent être envisagés (par exemple, les paramètres pour décrire ce qu'est un nom de famille ou une adresse postale), avec comme problématique sous-jacente la découverte automatique de la combinaison de paramètres permettant l'obtention de performances optimales, et (ii) si le programme doit évoluer ou s'adapter au cours du temps à de nouveaux paramètres ou à de nouvelles données qui lui seraient présentés.

Les algorithmes d'apprentissage permettent donc de s'affranchir de ces deux contraintes. Comme le rappelle [Wisniewski, 2007], « *plutôt que d'écrire une spécification formelle du comportement du programme, le programmeur fournit une base d'apprentissage composée d'exemples d'entrée et leur sortie attendue* ». L'expert fournit ainsi à la machine autant de caractéristiques qu'il le souhaite et s'affranchit de la première contrainte en déléguant à l'outil le choix des caractéristiques à utiliser. D'autre part, l'algorithme gère automatiquement les ajouts de caractéristiques ou les nouvelles données fournies par l'utilisateur pour peu qu'on lui fournisse des corpus annotés appropriés.

Avantages et inconvénients

Les méthodes à base d'apprentissage permettent de traiter rapidement de gros volumes de données. L'efficacité des modèles construits dans ces méthodes tient dans l'équilibre du rapport signal/bruit, les méthodes étant conçues pour être robustes.

15. « *Many approaches to traditional NER use dictionaries and gazetteers of person, organization, and location names. Ambiguous and out-of-vocabulary PHI reduce the contribution of dictionaries and gazetteers to de-identification and emphasize the value of studying context and language.* »

Ainsi, un système par apprentissage statistique produira des résultats de moins bonne qualité sur les informations peu annotées dans le corpus d'apprentissage.

Les inconvénients de ce type de méthode concernent l'absolue nécessité de disposer d'un corpus annoté qui couvre, pour chaque catégorie d'information, le maximum de variantes, de manière à construire des modèles qui soient les plus robustes possibles. Un grand nombre de variantes est ainsi nécessaire, aussi bien pour le contexte dans lequel évoluent les expressions à traiter *(contextes gauches et droits de l'expression)* que pour la constitution des expressions à traiter *(de manière à utiliser les caractéristiques issues des propriétés internes de ces expressions)*. Le choix des caractéristiques à étudier pour construire les modèles demeure complexe et l'utilisation qu'en fait le CRF n'est pas aisément compréhensible par l'expert. Enfin, la difficile compréhension des raisons qui ont poussé le modèle à produire certaines erreurs ne facilite pas l'amélioration des modèles ainsi construits.

2.3.2 Formalismes

On distingue deux types de modèles en matière d'apprentissage : les modèles génératifs et les modèles discriminants. Dans chaque formalisme, toutes les instances sont indépendantes les unes des autres. Dans ce travail de thèse, nous ne considérerons que les approches à base d'apprentissage supervisé.

Les modèles génératifs *(réseaux bayésiens naïfs, HMM)* modélisent la probabilité jointe $P(y, x)$, c.-à-d. la probabilité d'une étiquette y d'après un vecteur de caractéristiques x. Afin de prendre une décision sur l'étiquette à apposer sur un token, il est nécessaire de modéliser la probabilité conditionnelle $P(y|x)$. Le modèle génératif doit alors calculer la probabilité d'un vecteur de caractéristiques $P(x)$. Cette dernière opération se révèle particulièrement complexe à mener. Nous donnons dans l'équation 2.1 la manière dont la modélisation de la probabilité conditionnelle est réalisée dans le cadre des modèles génératifs.

$$P(y|x) = \frac{P(y, x)}{P(x)} \tag{2.1}$$

Les modèles discriminants *(régression logistique, entropie maximale, CRF de chaînes linéaires)* modélisent directement la probabilité conditionnelle $P(y|x)$ qui permet la prise de décision, c.-à-d. le choix de la séquence d'étiquettes à utiliser pour une séquence donnée de tokens.

Présentation

Les formalismes utilisés en traitement automatique des langues sont généralement des classifieurs linéaires ou log linéaires. Ils reposent sur des modèles discriminants adaptés à des étiquetages en séquences. Ces formalismes effectuent un apprentissage des relations qui existent entre un label *(nom, prénom, date, etc., « out » pour l'absence de label)* et des observations pour produire un modèle. L'application du modèle ainsi créé permet de prédire quelle est la séquence de labels la plus probable pour une séquence de tokens non étiquetés.

Ainsi, pour une séquence donnée de mots sous la forme d'un vecteur de caractéristiques *(casse du token, présence du token dans une liste, nombre de caractères, etc.)* $\mathrm{x} = (x_1, ...x_i, ...x_n)$, les modèles d'apprentissage vont chercher à identifier quelle est la séquence de labels la plus probable $\mathrm{y} = (y_1, ...y_i, ...y_n)$ qui maximise la combinaison des caractéristiques et de leurs poids.

La première décision concerne le type de tokénisation à effectuer. La tokénisation d'un texte consiste à segmenter le texte en « tokens », un token consistant en une suite de caractères *(tout type de caractères : alphabétiques, numériques, ponctuation)* comprise entre deux espaces. Deux types de tokénisation s'opposent : soit (i) une tokénisation simplement effectuée sur l'espace *(le texte est découpé en tokens autour de chaque espace typographique après avoir ajouté des espaces autour de chaque signe de ponctuation)*, soit (ii) une tokénisation réalisée de manière améliorée en préservant les abréviations et les décimales dans les nombres, donc en limitant l'ajout d'espaces autour des signes de ponctuation aux seules ponctuations du texte (ni la virgule dans les nombres décimaux, ni les points dans les acronymes et abréviations). La deuxième décision se rapporte au choix des caractéristiques à utiliser dans la construction du modèle (voir ci-après).

Plutôt que de construire un modèle par classe *(un modèle pour les noms, un autre pour les prénoms, etc.)* dont l'application produira des prises de décision incompatibles *(deux étiquettes différentes pour un même token)*, nous plaçons notre travail dans le cadre d'un modèle global *(multi-classe)* qui est l'approche suivie par les CRF de chaînes linéaires.

Caractéristiques

L'obtention de résultats de qualité est fortement dépendant des caractéristiques utilisées pour construire les modèles. Elles consistent en des éléments d'information qui seront associés à chaque token du document. Au niveau de l'apprentissage statistique, on distingue plusieurs catégories de caractéristiques.

Caractéristiques de surface. Des caractéristiques de surface, dites « lexicales », en inférant des propriétés à partir du token, propriétés traduites sous la forme de caractéristiques : la casse du token, la présence de ponctuation, la longueur du mot, la présence de caractères spéciaux, la présence de nombres, une désinence particulière, etc. Ces caractéristiques portent, soit uniquement sur le mot étudié, soit en tenant compte des voisins (bi- ou trigrammes).

Caractéristiques profondes. Des caractéristiques profondes, dites « riches », qui recouvrent plusieurs types du point de vue linguistique :

- des caractéristiques morpho-syntaxiques *(étiquetage en parties du discours)* ;
- des caractéristiques syntaxiques sur le mot et ses voisins, éventuellement en projetant sur les tokens la sortie d'une grammaire de dépendance ;
- et des caractéristiques sémantiques *(par exemple, les types sémantiques de l'UMLS, la présence du token dans une liste de noms, de prénoms, etc.)*.

Caractéristiques externes. Des caractéristiques externes :

- la position du token dans le document *(en rattachant le token à la section sous laquelle il apparaît)*. [Hara, 2006] a cependant indiqué que le gain était limité ;
- la fréquence globale du token dans le document ;
- l'indication de l'identifiant numérique du *cluster* auquel appartient le token, sur la base d'un *clustering* non supervisé. [16]

16. On appelle « cluster » un groupe de tokens. Le *clustering* non supervisé consiste à déléguer à la machine le processus de regroupement de tokens sur la base de propriétés partagées par les différents tokens, notamment fondées sur le voisinage dans lequel s'inscrit chaque token. Aucune indication n'est

Typologie

Il existe plusieurs types de classifieurs utilisables en traitement automatique des langues, parmi lesquels les arbres de décision, les champs aléatoires conditionnels, et les séparateurs à vaste marge.

Les arbres de décision

Les « arbres de décision » constituent une famille de modèles utilisés pour faire de la classification de données [Manning et Schütze, 2000]. Par analogie avec un arbre naturel, un arbre de décision se compose de nœuds, desquels partent des branches, jusqu'à atteindre des feuilles (parties terminales de l'arbre). Chaque nœud correspond à une question ou à une décision à prendre. En fonction de la réponse apportée à cette question ou de la décision prise, le parcours dans l'abre passera par l'une ou l'autre des branches partant de ce nœud. Le parcours dans l'arbre s'arrête dès qu'une feuille est atteinte.

Pour résoudre la problématique posée par le challenge i2b2 2006, [Szarvas et al., 2007] ont modifié un système de reconnaissance d'entités nommées pour traiter l'anonymisation. Leur système repose sur des arbres de décision qui ont été entraînés sur les entêtes des documents. Les caractéristiques utilisées par les auteurs pour construire les arbres de décision reposent sur : (i) des caractéristiques orthographiques *(capitalisation, longueur du mot)*, (ii) la fréquence du terme, (iii) la classe du mot précédent, (iv) le déclencheur présent dans l'expression, (v) des listes *(noms de famille, lieux)*, et (vi) des informations contextuelles *(position dans la phrase, entête de section le plus proche)*.

Les séparateurs à vaste marge

Le principe de base des séparateurs à vaste marge (SVM) repose sur le fait que l'algorithme va chercher à calculer l'hyperplan qui sépare le mieux un espace en classes [Vapnik, 1995] (figure 2.4). L'algorithme intègre généralement une fonction *kernel* (le noyau), qui permet de transposer un espace de données dans un autre espace qui sera linéairement séparable. Une optimisation est alors réalisée pour déterminer l'hyperplan, autrement dit, pour « *maximiser la distance entre l'hyperplan séparateur et les points les plus proches de chaque classe* » [Wisniewski, 2007]. Ainsi, les données seront les plus éloignées possible de l'hyperplan. Les systèmes linéaires *(liblinear)* conviennent aux cas où le nombre d'attributs est important par rapport au nombre d'instances.

Lors de sa participation au challenge i2b2 2006, [Hara, 2006] a mis au point une chaîne de traitements reposant sur quatre étapes : (i) des patrons syntaxiques ont été définis pour repérer les entêtes du document, potentiellement porteurs d'informations à anonymiser, (ii) des expressions régulières ont été produites pour traiter les informations numériques, puis (iii) un classifieur a été créé pour classer les phrases selon qu'elles contiennent ou non des informations à anonymiser, enfin, (iv) un chunker fondé sur SVM a été mis en place pour identifier et typer les informations à anonymiser. L'auteur indique cependant que le repérage des entêtes de document, pour le corpus i2b2 2006, n'a apporté qu'un gain très limité dans la détection des in-

fournie à la machine, d'où le caractère « non supervisé » de ce regroupement.

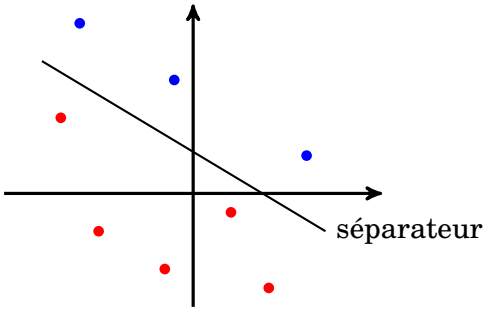


FIGURE 2.4 – Séparateur placé à la distance maximale des individus des deux classes

formations devant être anonymisées. Il précise d'autre part que son meilleur résultat est celui qui n'utilise pas l'étape de classification des phrases.

Une autre équipe participant au défi i2b2 2006 a utilisé un classifieur à base de SVM [Guo et al., 2006]. Dans un premier temps, les auteurs ont utilisé le système d'extraction d'information ANNIE de la plate-forme GATE [17] en complément du classifieur SVMlight, [18] ce dernier ayant la particularité d'être directement utilisable depuis la plate-forme GATE. Les auteurs rapportent avoir ajouté des caractéristiques supplémentaires sur chaque token du corpus, en particulier pour traiter plus efficacement les noms, hôpitaux, âges et lieux, mais également pour prendre en compte les différents formats de date et de numéros de téléphone. Enfin, le classifieur SVMlight a été utilisé pour la classification finale de chaque token.

En dehors d'une participation classique au challenge i2b2 2006, les organisateurs du challenge ont mis au point un système d'anonymisation nommé « Stat De-id » qui a été évalué sur les données du challenge [Uzuner et al., 2008]. Ce système repose sur l'utilisation de deux outils : (i) l'outil LibSVM [19] [Chang et Lin, 2011] pour classer les tokens dans une ou plusieurs catégories d'élément anonymisé, voire dans aucune catégorie dans le cas où le token doit rester en clair dans le document, et (ii) le parser Link Grammar [20] de manière à spécifier la phrase dans laquelle apparaît le token. Les auteurs rapportent des taux de rappel (0,97) et de précision (0,99) très élevés.

Dans le détail, les auteurs ont mobilisé trois types de caractéristiques pour produire le vecteur de caractéristiques utilisé pour construire leur modèle.

Sur le plan lexical, des caractéristiques de base, telles qu'on peut les retrouver dans un grand nombre d'outils à base d'apprentissage, ont ainsi été utilisées : (i) le mot cible *(tous les mots du corpus sont intégrés au vecteur)*, (ii) des bigrammes lexicaux *(les deux mots qui précèdent, les deux mots qui suivent)*, (iii) la capitalisation *(le token commence-t-il par une capitale ?)*, (iv) la ponctuation *(présence/absence de tiret et de barre oblique)*, (v) les nombres *(présence/absence de nombres dans le token)*, et (vi) la longueur *(nombre de caractères dans le mot)*.

Des caractéristiques d'ordre syntaxique ont également été mobilisées, avec : de manière assez traditionnelle, (i) les étiquettes morpho-syntaxiques du token et des

17. http://gate.ac.uk, University of Sheffield, Sheffield, Royaume-Uni. Plateforme pour le traitement automatique des langues.

18. http://svmlight.joachims.org, University of Dortmund, Allemagne. Outil de classification automatique fondé sur le formalisme des SVM.

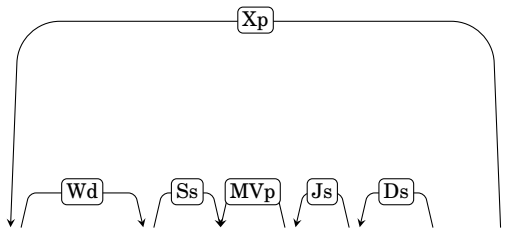
19. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, National Taiwan University, Taiwan. Outil de classification automatique reposant sur le formalisme des SVM.

20. http://www.link.cs.cmu.edu/link/, Carnegie Mellon University, Pittsburgh, PA. Outil d'analyse syntaxique de l'anglais fondé sur les structures syntaxiques des éléments de la phrase.

deux tokens qui suivent et qui précèdent, et de manière plus complexe et semble-t-il plus utile (ii) l'utilisation de bigrammes syntaxiques. Le Link Grammar Parser a ainsi été utilisé pour effectuer un rattachement, au token étudié, des deux termes qui lui sont liés à gauche et à droite. Un lien syntaxique est d'abord établi entre chaque token de la phrase (exemple 11.a), puis des bigrammes syntaxiques gauches et droits sont établis (exemple 11.b pour les bigrammes produits pour le token *lives*), chaque bigramme intégrant deux tokens accompagnés de l'étiquette [21] qui caractérise leur relation : (a) avec le token suivant pour les bigrammes gauches ou (b) avec le token précédent pour les bigrammes droits.

(11)

a.



b. {LEFT-WALL-Wd, John-Ss} lives {with-MVp, brother-Ds}

Enfin, sur le plan sémantique, trois informations ont été utilisées : (i) l'identifiant du syntagme nominal qui contient le mot cible dans le MeSH, (ii) la présence ou l'absence du token — et des deux qui précèdent et qui suivent — dans une liste de lieux, hôpitaux et noms, et (iii) l'entête de section.

Les champs aléatoires conditionnels

Le formalisme des champs aléatoires conditionnels (CRF) [Lafferty et al., 2001, Sutton et McCallum, 2006] est un modèle graphique linéaire non dirigé. Les CRF permettent facilement de prendre en compte le contexte pour étiqueter une séquence de tokens. [McCallum, 2003] a notamment démontré que l'efficacité des CRF est fonction de deux décisions à prendre.

Lors de leur participation au challenge i2b2 2006, [Aramaki et al., 2006] ont produit un système d'anonymisation reposant sur les CRF. Les auteurs ont d'abord manuellement ajouté une information sur chaque token du corpus d'apprentissage selon que le token est une donnée à protéger (PHI) ou non. L'approche suivie repose sur deux phases d'apprentissage. Une première phase permet d'apprendre les caractéristiques locales *(autour des mots)*, non locales *(longueur de la phrase, position dans le document)* et externes *(dictionnaires)*. Dans un deuxième temps, les caractéristiques précédentes sont utilisées en complément de quatre nouvelles caractéristiques pour attribuer le label.

Sur ce même challenge, [Wellner et al., 2007] ont adapté deux outils de REN *(Carafe, LingPipe)* qui font de l'étiquetage en séquence en considérant de manière habituelle qu'un token est au début, à l'intérieur ou à la fin d'une expression à anonymiser. Alors que Carafe implémente des CRF, en particulier pour des tâches d'identification d'expressions, LingPipe est constitué d'un ensemble de classes Java créées pour le

21. http://www.link.cs.cmu.edu/link/dict/summarize-links.html, « D » connecte un déterminant et un nom, « J » connecte une préposition et son objet, « S » connecte un nom sujet et un verbe, « W » connecte le sujet de la clause principale au mur gauche, enfin « MV » connecte verbes et adjectifs qui modifient l'expression qui suit.

traitement automatique des langues, parmi lesquelles figure un système de REN (fondé sur les chaînes de Markov cachées). L'utilisation conjointe de ces deux outils a permis aux auteurs de se classer premiers au challenge i2b2 2006.

[Gardner et Xiong, 2008] ont mis au point un système nommé « HIDE » pour anonymiser quatre catégories de termes : les noms, les âges, les dates, et les identifiants numériques. Ce système a été conçu comme un système de reconnaissance d'entités nommées et repose sur des CRF. Les auteurs précisent avoir utilisé les caractéristiques suivantes : le mot précédent, le mot suivant, la capitalisation du token, la présence d'un caractère spécial dans le token, et si le token est un nombre. Dans leurs expériences, les auteurs rapportent une exactitude de 0,982.

2.3.3 Construction du modèle

Les outils d'apprentissage que nous avons utilisés prennent en entrée deux éléments : (i) un fichier de configuration qui décrit la manière dont le modèle sera créé *(choix des caractéristiques, taille de la fenêtre, types de n-grammes de tokens et/ou de caractéristiques à utiliser, etc.)* et (ii) un corpus annoté *(annotation en caractéristiques et en exemples de sorties attendues)*.

Le corpus annoté se présente sous la forme d'un fichier tabulaire [22] (tableau 2.4) créé à l'issue d'une étape de tokénisation du fichier d'origine. Ce tabulaire intègre trois ensembles de colonnes : (i) la première colonne du tabulaire contiendra généralement le token issu du texte d'origine, (ii) la dernière colonne sera obligatoirement celle de la réponse attendue *(annotation de référence)*, et (iii) les colonnes intermédiaires seront celles des différentes caractéristiques utilisées pour décrire les propriétés des tokens *(casse typographique, étiquetage en parties du discours, etc.)*.

Token	Casse	Partie du discours	Annotation
Je	Mm	PRO:PER	O
revois	mm	VER:pres	O
avec	mm	PRP	O
plaisir	mm	NOM	O
madame	mm	NOM	O
Valérie	Mm	NAM	B-nom
Daumard	Mm	NAM	I-nom

TABLE 2.4 – Extrait d'un tabulaire : quatre colonnes séparées par une tabulation

Schémas d'annotation : BIO vs. BILOU

Le schéma d'annotation correspond au mode de représentation des données utilisé dans le fichier tabulaire, notamment dans la colonne de la réponse attendue. S'il est d'usage d'utiliser ce schéma d'annotation dans la colonne de réponse attendue, ce schéma peut également être utilisé dans les colonnes de caractéristiques. Le schéma le plus connu et le plus couramment utilisé est le schéma BIO. Depuis peu, une extension de ce schéma a été proposée sous le nom de BILOU (voir tableau 2.5).

Le schéma BIO. Ce schéma comprend trois indices qui permettent de renseigner la position d'un token dans l'expression annotée :

22. On appelle « fichier tabulaire » un fichier dont les colonnes sont séparées par une tabulation.

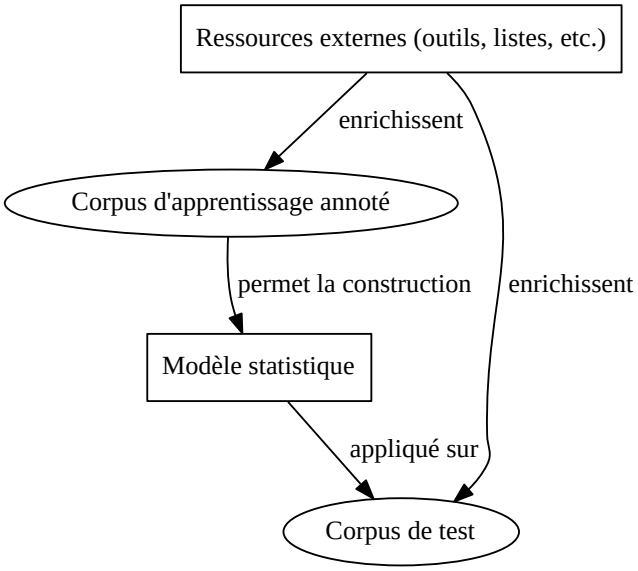


FIGURE 2.5 – Architecture d'un système par apprentissage supervisé

– B *(begin)* : cet indice permet de noter que le token est au début de la portion annotée. Dans le cas d'une annotation portant sur un seul token, celui-ci sera également le dernier de la portion. On n'indiquera cependant pas le fait qu'il est le dernier.
– I *(in)* : cet indice permet de noter que le token est à l'intérieur de la portion annotée. L'indication de rang n'est pas davantage précisée. L'indice « I » sera donc utilisé pour noter les tokens, du deuxième jusqu'au dernier dans la portion.
– O *(out)* : cet indice permet de noter que le token est en dehors de toute portion annotée.

Le schéma BILOU. Puisque le schéma BIO ne permet pas de noter le dernier token d'une portion, ni la position isolée d'un token annoté, une extension de ce schéma a été proposée sous le nom BILOU [Ratinov et Roth, 2009], également présent dans la littérature sous le nom BWEMO *(Begin, Whole, End, Middle, Out)*. Aux trois indices précédents s'ajoutent deux nouveaux indices qui réduisent les cas d'application des trois premiers indices :

– B *(begin)* : l'indice sert uniquement à marquer le premier token d'une portion annotée composée d'au-moins deux tokens.
– I *(in)* ou M *(middle)* : l'incide sert uniquement à noter les tokens qui ne sont ni le premier, ni le dernier de la portion annotée.
– L *(last)* ou E *(end)* : cet indice permet de noter que le token est le dernier de la portion annotée. Il n'est utilisé que dans le cas d'annotation portant sur plusieurs tokens.
– O *(out)* : comme dans le précédent schéma, sert à noter les tokens situés en dehors de toute portion annotée.

– U *(unit)* ou W *(whole)* : cet indice permet de noter que l'annotation ne porte que sur un seul token, qui est alors isolé.

Nous donnons dans le tableau 2.5 un exemple d'annotation des sorties attendues pour une phrase dans les deux schémas d'annotation. Les catégories de sorties attendues sont celles que nous avons utilisées dans nos expériences d'anonymisation.

Token	Schéma BIO	Schéma BILOU
Monsieur	O	O
Théodore	B-prénom	U-prénom
Bauche	B-nom	U-nom
(	O	O
24	B-date	B-date
/	I-date	I-date
06	I-date	I-date
/	I-date	I-date
46	I-date	L-date
)	O	O
est	O	O
malheureusement	O	O
revenu	O	O
dans	O	O
le	O	O
service	O	O
du	O	O
28	B-date	B-date
avril	I-date	I-date
au	I-date	I-date
5	I-date	I-date
mai	I-date	I-date
1993	I-date	L-date
pour	O	O
la	O	O
constitution	O	O
d'	O	O
un	O	O
nouvel	O	O
infarctus	O	O

TABLE 2.5 – Différences d'annotations entre les schémas BIO et BILOU

Privilégier une métrique : le biais du rappel

Introduction. En matière d'évaluation des résultats produits par un anonymiseur, notamment en domaine médical, on privilégie en règle générale le rappel sur la précision. Autrement dit, on préférera disposer d'un système qui anonymise toutes les informations devant l'être, permettant l'obtention d'un rappel le plus élevé possible, quitte à tolérer une sur-anonymisation, conduisant à une baisse de la précision. Lors

de l'évaluation proprement dite, il est possible de faire varier la valeur de β dans la F-mesure (voir section 3.2.2), de manière à privilégier le rappel sur la précision. Ce faisant, seule change la valeur finale de la F-mesure, tandis que les taux de rappel et de précision ne changent pas.

Méthodes. Depuis quelques années, les méthodes à base d'apprentissage évoluent de manière à favoriser, lors de l'application d'un modèle, une mesure plutôt qu'une autre. C'est dans cette optique qu'a émergé la problématique de la récompense d'une métrique choisie par l'utilisateur. [Culotta et McCallum, 2004] ont suggéré plusieurs moyens de générer une estimation de la confiance.

Une alternative à ces suggestions a été proposée par [Minkov et al., 2006]. L'algorithme proposé permet, soit de biaiser le rappel, de manière à augmenter la valeur finale du rappel, soit de biaiser la précision, pour augmenter la valeur finale de la précision. L'algorithme repose sur le principe d'attribution de poids dans les annotations. Ainsi, le moyen proposé pour biaiser le rappel consiste à étudier le poids affecté aux annotations portant la valeur « O » pour décider de changer la valeur d'origine « O » en « B » ou « I », et réciproquement pour biaiser la précision. Ainsi, l'étiquetage sera forcé *(de « O » vers « B » ou « I »)* en cas de poids fortement négatif tandis que l'étiquetage sera supprimé *(de « B » ou « I » vers « O »)* avec un poids fortement positif.

Une autre méthode consiste à enchaîner deux systèmes en cascade, le premier ayant pour objectif de privilégier le rappel quelles que soient les conséquences sur la précision, et le second système visant à améliorer la précision globale au moyen de règles de post-traitements. C'est notamment l'approche suivie par [Ferrández et al., 2012] que nous détaillons en section 2.4.3.

Expérimentations. Des expériences de biais du rappel ont été réalisées par [Deléger et al., 2013] sur un corpus de 3 503 documents cliniques provenant du CCHMC [23] au moyen de deux outils existants reposant sur le formalisme des champs aléatoires conditionnels Mallet [24] [McCallum, 2002] et MIST [25] [Wellner et al., 2007]. Les auteurs ont également appliqué une fonction de *tuning* via l'algorithme de [Minkov et al., 2006], de manière à biaiser le rappel et augmenter sensiblement la valeur de cette mesure. La recherche des meilleurs paramètres pour chaque outil pour biaiser le rappel a permis d'augmenter sensiblement les valeurs du rappel (de 0,9192 à 0,9366 avec Mallet et de 0,9281 à 0,9358 avec MIST). Cependant, puisque le biais du rappel consiste à augmenter la taille des portions anonymisées, cette technique engendre une sur-anonymisation importante. De ce fait, la valeur de la précision décroît plus rapidement que n'augmente celle du rappel, conduisant également à une baisse globale de la F-mesure (voir tableau 2.6 et figure 2.6).

Système	*Version sans biais*			*Version avec biais*		
	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
Mallet	0,9192	**0,9508**	**0,9348**	**0,9366**	0,9151	0,9257
MIST	0,9281	**0,9279**	**0,9280**	**0,9358**	0,8803	0,9072

TABLE 2.6 – Évaluation des anonymisations produites en fonction du biais du rappel

23. http://www.cincinnatichildrens.org/, Cincinnati Children's Hospital Medical Center, Cincinnati, OH.
24. http://mallet.cs.umass.edu/, Mallet : MAchine Learning for LanguagE Toolkit.
25. http://mist-deid.sourceforge.net/, MIST : MITRE Identification Scrubber Toolkit.

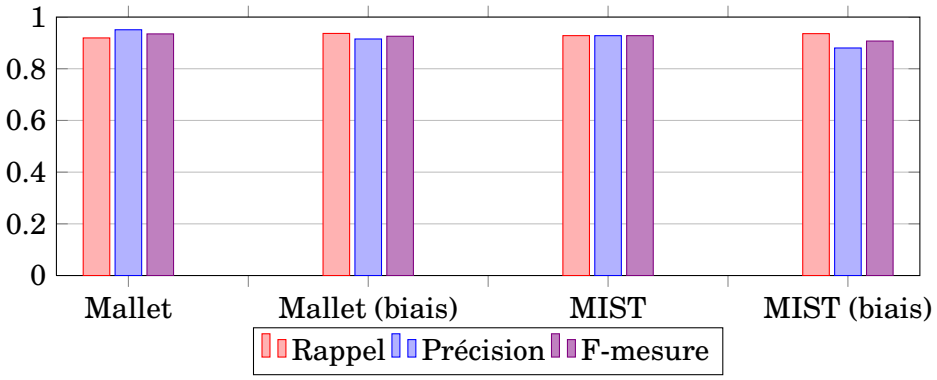


FIGURE 2.6 – Évaluation des anonymisations produites par Mallet et MIST en fonction du biais du rappel

De ces expériences, on peut donc conclure que le biais du rappel permet d'augmenter sensiblement la couverture des informations devant être anonymisées, mais dans le même temps, cela conduit à dégrader rapidement la précision de l'anonymisation.

Conséquences. [Deléger et al., 2013] ont également cherché a étudier l'impact de l'anonymisation sur une tâche d'extraction d'information *(une tâche d'extraction des noms de médicament)*. Sur cette tâche, la F-mesure se monte à 0,9264 sur le corpus non anonymisé. Puisque le biais du rappel engendre une sur-anonymisation, il demeure moins d'informations en clair, et les performances de l'extraction décroissent légèrement (MIST avec un biais de -3 : F=0,9253 vs. 0,9282 sans biais ; Mallet avec un seuil de 0,93 : F=0,9266 vs. 0,9278 sans biais).

Sur la base de ces différentes expériences, nous pouvons donc conclure que l'utilisation de techniques permettant de biaiser le rappel permet effectivement d'augmenter sensiblement la valeur du rappel, mais la qualité globale du corpus s'en trouve plus fortement dégradée et, conséquence importante, ne facilite pas les traitements ultérieurs appliqués sur le corpus ayant fait l'objet de ce type de traitement. Dans notre travail d'anonymisation automatique, nous n'envisageons donc pas de recourir à ces méthodes de biais du rappel.

2.3.4 Faire abstraction du paramétrage

Partant du principe que les approches symboliques souffrent d'un temps de développement, de maintenance et d'extension considérable, et que les méthodes par apprentissage nécessitent un volume de données annotées et une connaissance aiguë des outils pour la sélection des caractéristiques, [Aberdeen et al., 2010] ont développé MIST,[26] un outil d'aide à l'anonymisation. Le principe de cet outil consiste à offrir à l'utilisateur, sous un même environnement de travail, une interface d'annotation de documents et un outil d'apprentissage statistique fondé sur Carafe [Wellner, 2009] et les champs aléatoires conditionnels. Initialement conçu pour la participation à l'édition 2006 du défi i2b2, cet outil peut facilement être adapté à de nouvelles classes *(les auteurs ont ajouté de nouvelles classes depuis le défi i2b2)* et de nouvelles données *(l'outil a été appliqué sur un corpus de documents cliniques de l'institut Vanderbilt)*.

26. http://mist-deid.sourceforge.net/, MIST, the MITRE Identification Scrubber Toolkit.

Le fonctionnement global de MIST repose sur quatre étapes : (i) l'utilisateur annote un premier jeu de données *(via une interface web d'annotation des données)*, et (ii) lance l'apprentissage sur la base de ces annotations *(l'outil apprend alors contextuellement les caractéristiques à utiliser et crée autant de modèles qu'il y a de catégories à traiter)*, puis (iii) il corrige les sorties annotées et (iv) relance le processus d'annotation et d'apprentissage jusqu'à atteindre les performances voulues.

Il est particulièrement intéressant de constater que selon le type d'information traitée *(informations numériques vs. nominatives)*, le nombre d'exemples annotés permettant l'obtention de résultats de qualité n'est pas le même. Ainsi, il est possible d'obtenir une F-mesure supérieure à 0,95 pour les données numériques *(âges, identifiants numériques)* avec moins d'une centaine d'exemples annotés, alors qu'il faudra autour de mille exemples annotés pour les données nominatives *(noms)* ou combinant plusieurs éléments *(dates)*. D'autre part, la qualité de l'anonymisation est également largement dépendante du type de document sur lequel est appliquée l'anonymisation ; les auteurs rapportent ainsi des F-mesures différentes, bien que cependant élevées, selon le type de document traité : 0,996 sur les comptes rendus hospitaliers, 0,996 sur les récapitulatifs de prescription, 0,943 pour les lettres de suivi et 0,934 pour les résultats de laboratoire.

2.4 Les méthodes hybrides

L'hybridation des caractéristiques issues des deux approches précédentes semble donner de meilleurs résultats. Ainsi, la mobilisation de ressources linguistiques pour construire les modèles [Uzuner et al., 2008] semble surpasser les approches précédentes et propose une qualité d'anonymisation proche de la référence. D'autres méthodes combinatoires peuvent être envisagées, notamment en utilisant conjointement les deux approches et en délégant le choix des informations à anonymiser à un processus de vote automatique.

2.4.1 Le symbolique pour produire les caractéristiques de l'apprentissage

Principe. Le premier type d'hybridation consiste à utiliser les méthodes à base de règles pour produire les tabulaires de caractéristiques utilisés pour créer le modèle (voir figure 2.7). [Meystre et al., 2010] rapportent que ce type d'hybridation a permis aux systèmes qui l'ont utilisé d'obtenir les meilleurs résultats lors de l'édition 2006 du challenge i2b2.

Indices internes et externes. Il existe deux types d'indices pour effectuer l'anonymisation de certains types de contenus dans des documents cliniques, selon que l'on fonde l'analyse sur l'élément à anonymiser ou sur l'entourage de cet élément. Travaillant sur l'identification et la catégorisation sémantique des noms propres, [McDonald, 1993] a mis en évidence l'intérêt d'utiliser les indices internes et externes de ces mots. Les indices internes renvoient aux éléments compris dans la séquence de mots formant le nom, notamment les déclencheurs *(abréviation du type juridique de l'entreprise — SA, SARL, etc. — ou prénom connu devant le nom de famille)*. À l'opposé, les indices externes concernent le contexte dans lequel occurrent les noms à traiter : par exemple, des déclencheurs non inclus dans le nom *(Monsieur X, l'entreprise X)* ou des verbes typiques d'un nom *(X a déclaré, X s'est rendu)*.

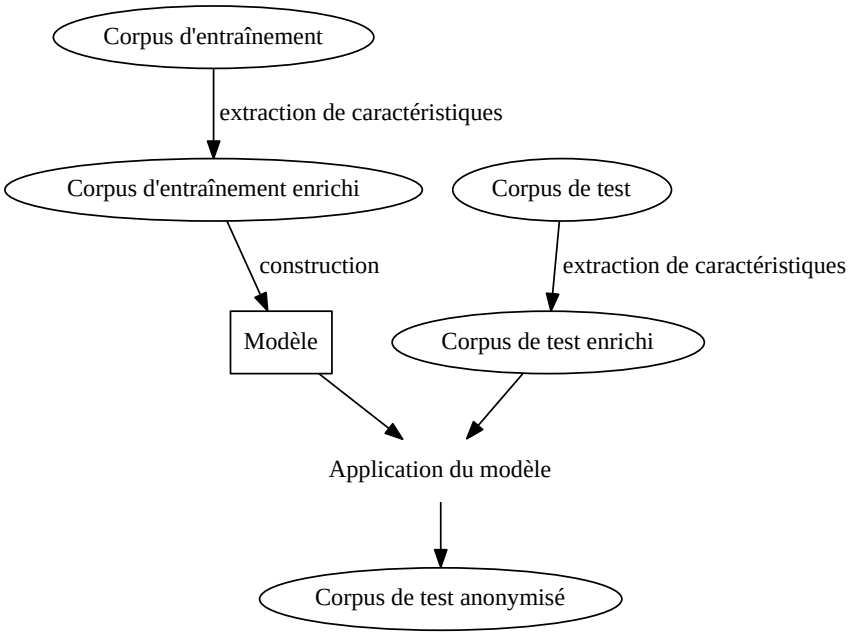


FIGURE 2.7 – Utilisation des règles pour produire les caractéristiques

S'appuyer sur le contexte d'apparition de ces informations reprend le principe selon lequel des informations qui occurrent dans des contextes similaires partagent des propriétés communes, qui est à rapprocher de l'hypothèse distributionnelle établie par [Harris, 1985]. Appliquée à l'anonymisation, cette hypothèse voudrait que des informations à anonymiser d'un même type se retrouvent dans des contextes équivalents. L'avantage des indices externes est leur robustesse, puisqu'ils sont insensibles aux erreurs de frappe et aux variantes orthographiques ou syntaxiques internes affectant les informations à traiter. L'exploitation des caractéristiques linguistiques et de la spécificité de formulation de ces indices peut se faire au moyen d'approches à base d'apprentissage.

Expérimentations. Dans le challenge i2b2 2006, [Uzuner et al., 2007] distinguent trois types d'indices qui ont été utilisés pour produire les caractéristiques nécessaires à la création des modèles : (i) des indices globaux, qui correspondent aux informations relatives au document — voire au corpus — dans lequel apparaît le token *(position du token dans la phrase, longueur de la phrase, mots de la phrase précédente, label majoritairement attribué au token dans le document/corpus, degré de confiance sur le label PHI, information de l'entête)*, (ii) des indices locaux, qui correspondent aux caractéristiques internes directement inférées des tokens *(n-grammes d'informations lexicales, tokens, n-grammes de tokens, information phrastique, étiquetage en partie du discours, information orthographique, taille du token, affixes, caractères particuliers, déclencheurs, n-gramme ou arbre de dépendance, fréquence)*, et (iii) l'utilisation de ressources externes *(présence du token dans un dictionnaire)*.

Discussion. Cet inventaire d'indices utilisés pour créer les fichiers de caractéristiques ont été utilisés de manière variée par chaque système. Aucun n'intègre l'ensemble de ces caractéristiques. Il est à noter que le système qui a utilisé le plus de caractéristiques différentes *(le système à base de CRF de l'équipe Aramaki)* s'est classé troisième. Avec le même formalisme *(CRF)* et moins de caractéristiques, le système de l'équipe Wellner s'est mieux classé au niveau de la F-mesure (0,9806 contre 0,9697) ; bien que plus précis, le système Aramaki n'a pas permis d'obtenir un rappel supérieur à 0,9494. Le système Wellner semble avoir privilégié les bigrammes de caractéristiques lexicales et des affixes, soit une importance accrue accordée aux informations sémantiques.

2.4.2 Le symbolique en pré- et post-traitements de l'apprentissage

Principe. Un deuxième type d'hybridation assez simple à mettre en œuvre consiste à utiliser les méthodes symboliques pour ajouter des pré-traitements sur les données d'origine et des post-traitements sur les données produites par les méthodes par apprentissage (voir figure 2.8).

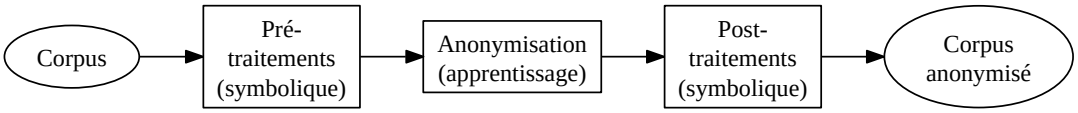


FIGURE 2.8 – Ajout de pré- et post-traitements autour de l'apprentissage

Expérimentations. En matière d'hybridation des traitements, [Deléger et al., 2013] ont utilisé MIST et Mallet dans leur configuration de base *(caractéristiques et ressources par défaut)*, qu'ils ont complétés de nouvelles caractéristiques ainsi que des règles de pré- et post-traitements. Ce travail a été mené dans la perspective de mesurer l'impact de l'anonymisation sur les performances de systèmes d'extraction d'information. Appliqués sur le corpus CCHMC, les auteurs rapportent que les versions avec ajout de règles *(hybridation par ajout de pré- et post-traitements)* et de caractéristiques sont plus efficaces que dans leur version de base avec un gain d'un point de F-mesure, confirmant l'intérêt d'une hybridation des méthodes (voir tableau 2.7 et figure 2.9).

Système	*Version de base*			*Version avec règles*		
	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
Mallet	0,8986	**0,9525**	0,9248	**0,9192**	0,9508	**0,9348**
MIST	0,9102	0,9205	0,9154	**0,9281**	**0,9279**	**0,9280**

TABLE 2.7 – Évaluation des anonymisations produites avec ou sans traitement complémentaire

Conséquences sur les traitements futurs. Un autre point intéressant de cette étude concerne l'impact du type d'anonymisation effectuée sur les performances fu-

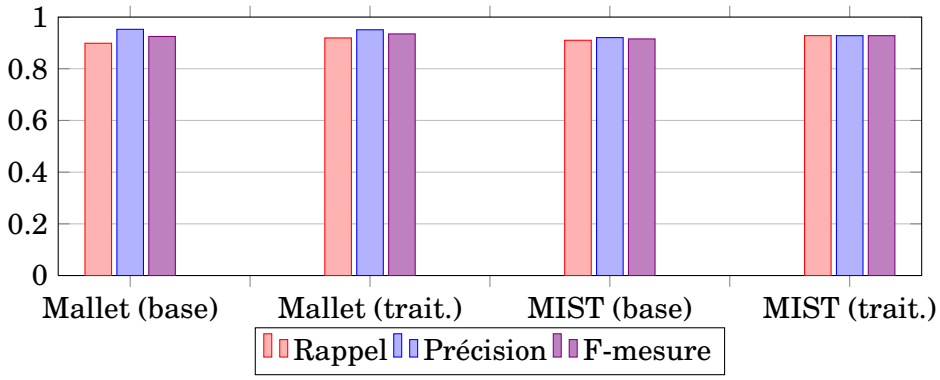


FIGURE 2.9 – Évaluation des anonymisations produites par Mallet et MIST avec ou sans traitement complémentaire

tures de tâches appliquées sur le corpus. Les auteurs ont choisi une tâche d'extraction des noms de médicaments pour illustrer cet impact. Ainsi, la F-mesure se monte à 0,9264 sur le corpus non anonymisé. On remarque également que l'anonymisation au moyen d'informations vraisemblables *(utilisation de pseudonymes)* dégrade légèrement les résultats (F=0,9254 avec Mallet) par rapport à la version du corpus où les informations auront été masquées par un caractère générique *(le caractère étoile « * »)* (F=0,9278 avec Mallet, F=0,9282 avec MIST, F=0,9283 avec une anonymisation manuelle).

2.4.3 Cascade de systèmes

Un dernier type d'hybridation repose sur l'utilisation des plusieurs systèmes utilisés en cascade, l'un après l'autre, pour effectuer l'anonymisation. Contrairement aux hybridations précédentes où l'anonymisation est réalisée par les méthodes par apprentissage statistique et dans lesquelles les méthodes symboliques servent pour les pré et post-traitements ou pour produire les caractéristiques, ce type d'hybridation revient, soit (i) à enchaîner plusieurs systèmes, soit (ii) à utiliser plusieurs systèmes en parallèle puis à fusionner les anonymisations effectuées.

Succession de systèmes

Principe. Dans le premier type de cascade, on dispose de deux systèmes que l'on applique successivement, le second prenant en entrée la sortie du premier système (voir figure 2.10). Dans ce type d'hybridation, le premier système agit sur les types d'information qu'il est en mesure de traiter, tandis que le deuxième complète les traitements en agissant sur les informations qui n'auront pas été traitées par le premier. Dans ce schéma, chaque système agit sur son domaine de spécialité, dans le sens d'une spécialisation des systèmes en fonction des types d'information pour lesquels chaque système a été conçu *(les entités numériques et les entités nominatives présentes dans des listes pour le système à base de méthodes symboliques).*

Expérimentations. À l'occasion de l'enchaînement de deux outils reposant sur des méthodes symboliques, d'abord le script « deid.pl » *(issu du PhysioToolkit*[27]*)* puis

27. http://www.physionet.org/physiotools/

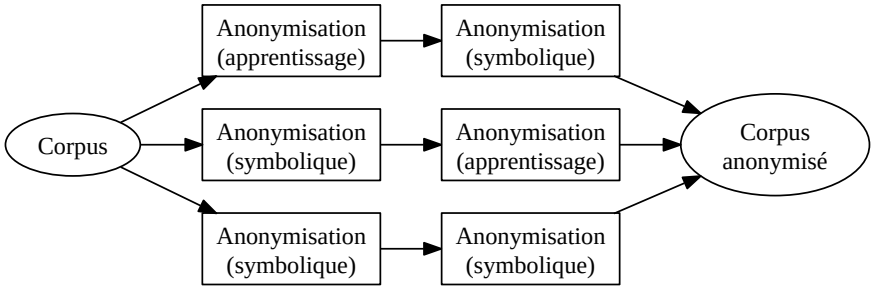


FIGURE 2.10 – Anonymisation par succession des différentes méthodes

le système MedLEE [28] [Friedman et al., 1994] *(utilisé pour repérer des entités nommées du domaine médical)*, une équipe a pu mettre en évidence le fait que les performances de MedLEE étaient améliorées par rapport à celles obtenues par MedLEE testé isolément, confirmant l'intérêt d'utiliser en premier les approches symboliques [Morrison et al., 2009b].

Ordre. La succession de plusieurs méthodes implique qu'un système prend en entrée le résultat de l'application du premier système. Autrement dit, le deuxième système travaillera sur la base de ce qui aura été mis en évidence par le premier système. L'inconvénient majeur de ce type d'hybridation repose sur le fait que le deuxième système doit être en mesure, soit (i) d'appliquer ses traitements sur des données qui seront déjà porteuses d'informations *(en cas d'annotation embarquée)*, soit (ii) de prendre en compte les premières annotations *(en cas d'annotation débarquée)*. [29]

Pour le cas où le système à base de règles serait appliqué en second, les règles devront avoir été conçues de telle sorte qu'elles puissent travailler à la fois sur des documents vides de toute annotation *(pour le cas d'une portion devant être anonymisée et qui n'aura pas été traitée par l'approche par apprentissage)*, comme sur des documents déjà annotés. La prise en compte de tous les cas de figure dans les règles tenant compte du contexte augmente la complexité de la production des règles dans la réalisation de ce système.

Inversement, pour le cas où le système par apprentissage serait appliqué après les méthodes symboliques, la présence d'annotations implique que le(s) modèle(s) appliqué(s) ai(en)t été construit(s) sur des données déjà porteuses d'annotations, ce qui augmente considérablement le nombre de contextes et réduit *de facto* la robustesse du système.

Discussion. Dans les deux cas, il apparaît que la robustesse des systèmes qui interviennent en second ne peut être garantie, du fait de l'augmentation des contextes à

28. http://www.cat.columbia.edu/medlee.htm, Medical Language Extraction and Encoding system, Columbia University, New York, NY.

29. On appelle « annotation embarquée » des annotations réalisées directement sur les données, par exemple au moyen de balises SGML encadrant les expressions annotées. A l'inverse, une « annotation débarquée » consiste à reporter dans un fichier annexe les annotations et à faire référence aux expressions sur lesquelles s'appliquent ces annotations au moyen de coordonnées *(numéro des tokens de début et de fin de l'expression, ou position des caractères de début et de fin)*.

prendre en compte *(contextes annotés et non annotés)*. Si les méthodes symboliques sont réputées plus précises, l'application de telles méthodes en second ne permettra que très modérément d'améliorer la précision. En ce qui concerne les méthodes par apprentissage, elles permettent d'obtenir un meilleur rappel. Leur application en deuxième ne permettra toutefois de n'augmenter que marginalement les taux de rappel et de précision.

Systèmes en parallèle

Principe. Dans le deuxième type de cascade, plusieurs systèmes seront appliqués en parallèle. Les différents systèmes utilisés peuvent relever uniquement des méthodes par apprentissage statistique *(en mobilisant plusieurs formalismes)*, ou bien combiner méthodes symboliques et par apprentissage statistique, en considérant que chaque type de méthode et chaque formalisme permet de traiter de manière optimale certaines catégories d'information. Une procédure de vote final est ensuite appliquée et permet de fusionner les anonymisations réalisées par chaque système, de manière à ne retenir que les meilleurs traitements de chaque système (voir figure 2.11). Lors de la fusion des annotations produites par les différents systèmes, la gestion des annotations qui se recouvrent peut se révéler complexe et implique de prendre une décision concernant les différences de frontières d'annotation.

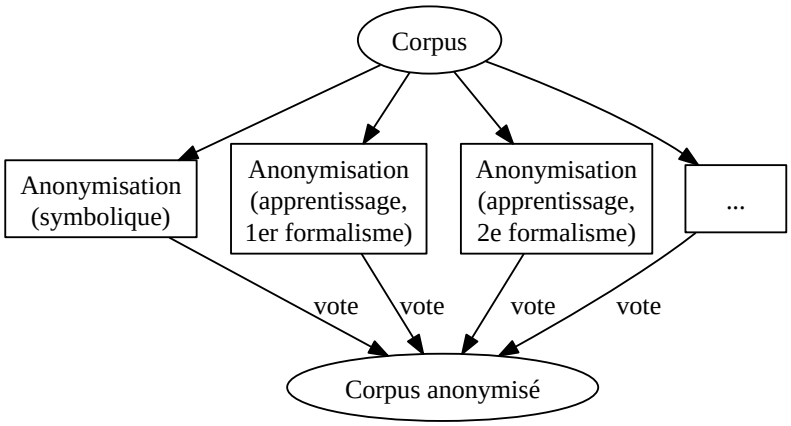


FIGURE 2.11 – Cascade de systèmes

Expérimentations.

HIDE. Le système HIDE [Gardner et Xiong, 2009] repose sur la combinaison de deux systèmes, un premier fondé sur le formalisme des CRF pour identifier et extraire les informations, et un deuxième système reposant sur les techniques à base de k-anonymat pour effectuer l'anonymisation finale. L'objectif de ce système en deux étapes consiste à préserver le maximum de données cliniques utiles tout en anonymisant les données sensibles.

L'approche suivie pour l'identification et l'extraction des informations à anonymiser repose sur quatre étapes principales : (i) une interface permet à l'utilisateur d'annoter les données qui doivent être anonymisées, de manière à fournir une base annotée pour l'apprentissage, (ii) un composant s'inspire des annotations précédentes pour générer les attributs à utiliser lors de la construction du modèle, (iii) un classifieur à base de CRF réalise une classification des éléments du texte, et (iv) un ensemble de post-traitements est alors appliqué pour nourrir de nouveau le classifieur, de manière à corriger la classification précédemment réalisée.

Sur la base de cette identification, le système d'anonymisation des données est alors appliqué. Ce système repose sur le modèle du k-anonymat [Sweeney, 2002] et de son extension, la l-diversité [Machanavajjhala et al., 2006].

Best of Breed. Dans cette perspective, [Ferrández et al., 2012] ont développé un système hybride nommé « *Best of Breed* » (BoB, « le meilleur de chaque type ») pour traiter les données cliniques du corpus Veteran's Health Administration (VHA). Ce système se compose de plusieurs processus montés en cascade sous la forme de composants UIMA.[30] Le traitement repose sur deux étapes principales, chacune bénéficiant de composants dédiés, dont le fonctionnement rappelle les algorithmes de biais du rappel présentés en section 2.3.3. La première étape consiste à privilégier le rappel, quelles que soient les conséquences sur la précision, au moyen (i) de patrons fondés sur des déclencheurs *(Mr, Dr)*, (ii) d'une projection de lexiques *(via le moteur d'indexation Lucène)* et (iii) le classifieur et le système de reconnaissance d'entités nommées du Stanford CoreNLP[31] et dont l'objectif consiste à prédire le format des informations à anonymiser. La seconde étape vise à améliorer la précision globale *(en filtrant le bruit généré par le précédent composant)*.

Ce système en deux étapes utilise plusieurs pré-traitements *(segmentation en phrases, tokénisation, étiquetage en parties du discours, segmentation en syntagmes, normalisation des mots)* à partir de composants OpenNLP[32] et cTAKES[33] [Savova et al., 2010]. Plusieurs classifieurs ont été créés à partir de LibSVM et LibLinear. Ils ont été entraînés à partir des sorties précédentes et permettent de décider si une annotation est correcte ou pas.

Trois configurations de caractéristiques ont été envisagées : (i) sans dictionnaire *(la configuration par défaut des outils)*, (ii) avec sélection des attributs à partir de trois listes *(dictionnaire de noms communs, villes américaines, noms de famille)*, mais les auteurs ont remarqué qu'il existe un problème pour entraîner HIDE simultané-

30. http://uima.apache.org, Unstructured Information Management Applications, IBM et Apache Software Foundation.

31. http://nlp.stanford.edu/software/corenlp.shtml, suite de systèmes de traitement automatique des langues, Stanford University, Stanford, CA. La suite intègre les composants suivants : étiquetage en parties du discours, reconnaissance d'entités nommées, classifieur CRF, etc.

32. http://opennlp.apache.org, plateforme de traitement automatique des langues fondée sur l'apprentissage statistique, Apache Software Foundation. L'outil intègre les composants suivants : tokénisation, segmentation en phrases, étiquetage en parties du discours, repérage d'entités nommées, segmentation en syntagmes, analyse syntaxique, résolution des coréférences.

33. https://wiki.nci.nih.gov/display/VKC/cTAKES+2.5, clinical Text Analysis and Knowledge Extraction System, Mayo Clinic, Boston, MA. Système d'extraction d'information depuis des documents cliniques fondé sur OpenNLP et l'architecture UIMA. L'outil intègre plusieurs composants parmi lesquels : segmentation en phrases, tokénisation à base de règles, normalisation, tokénisation d'après le contexte, étiquetage en parties du discours, segmentation en syntagmes, annotation par appariement à un dictionnaire, annotation d'après le contexte, détection de négation, analyse en dépendance, identification du statut de fumeur chez les patients, annotation des noms de médicaments.

ment sur l'ensemble des dictionnaires, et (iii) avec toutes les caractéristiques des dictionnaires.

L'évaluation a été réalisée au niveau des portions anonymisées et non au niveau des mots ; ce type d'évaluation se fonde donc sur la reconnaissance en même temps que sur le typage. Pour éviter une évaluation stricte *(trop dure)* et partielle *(risque d'oubli d'une partie d'un mot)* mais pour ne pas pénaliser des erreurs de frontières dues à des mots outils, l'évaluation a été faite sur la base d'un relâchement de contraintes *(dans cette perspective, une portion sera considérée comme correcte si elle intègre tous les éléments ou plus de la référence)*. Le calcul final se fait au moyen de la F_2-mesure (voir section 3.2.2).

Afin de comparer les performances de l'outil BoB, les auteurs ont appliqué les outils MIST et HIDE sur les mêmes corpus. Sur le corpus VHA, le système MIST a permis un meilleur traitement de l'anonymisation que le système HIDE. Le système BoB obtient cependant les meilleurs résultats. La configuration avec sélection d'attributs est meilleure que celle n'utilisant aucun dictionnaire et que celle utilisant, au contraire, toutes les caractéristiques.

Afin de tester la robustesse des systèmes, les auteurs ont appliqué les trois systèmes *(BoB, HIDE, MIST)* sur le corpus i2b2 2006, au moyen de deux expériences : (i) un entraînement et une évaluation sur le même corpus *(i2b2)*, et (ii) un entraînement sur le corpus VHA mais une évaluation effectuée sur le corpus i2b2. Cette fois-ci, le système HIDE obtient de meilleurs résultats que le système MIST. Cependant, le système des auteurs BoB continue d'obtenir les meilleurs résultats de tous les systèmes. Les auteurs notent toutefois que les résultats chutent si le système a été entraîné et appliqué sur deux corpus différents, ce qui correspond au comportement attendu d'un système par apprentissage confronté à des situations et des contextes qu'il n'aura pas vus lors de l'apprentissage.

2.5 Synthèse

En matière d'anonymisation automatique, plusieurs méthodes co-existent, les méthodes à base de règles d'une part, les méthodes par apprentissage statistique d'autre part, et l'hybridation de ces deux approches.

Méthodes à base de règles. Les méthodes à base de règles, dites « méthodes symboliques », consistent à utiliser des listes et des règles définies au moyen de connaissances d'experts. Si les résultats obtenus se révèlent de bonne qualité comparativement à d'autres approches, les méthodes symboliques souffrent de plusieurs inconvénients. Les coûts temporels et humains de production des règles et de maintenance sont assez élevés. D'autre part, l'utilisation de ce type de méthode sur un nouveau corpus ou un nouveau domaine médical suppose une adaptation des règles tout aussi coûteuse. L'utilisation d'informations externes en complément des règles, tels qu'un étiquetage en parties du discours ou une annotation sémantique au moyen du Metathesaurus de l'UMLS, permet d'améliorer sensiblement la qualité des traitements effectués. Sur les noms de substances actives, l'absence de liste peut être palliée par la formalisation des règles qui ont permis de nommer ces substances.

Un autre moyen de procéder à une anonymisation par le biais de méthodes à base de règles consiste à adapter un système de repérage d'entités nommées aux besoins de l'anonymisation. Le principe suivi par cette approche repose sur l'identification d'entités nommées comme étape préalable à l'anonymisation de ces entités. D'après

les expériences réalisées, l'adaptation d'un système de REN est rapide. Sur l'anglais, le système Carafe initialement conçu comme un système de repérage d'entités nommées a été adapté au challenge i2b2 2006. Sur le français, un autre système de REN a été adapté et complété de règles de post-traitements. Si l'adaptation semble rapide et facile à mettre en œuvre, les spécificités de la langue médicale et du corpus clinique ne permettent pas aux techniques de REN de traiter toutes les catégories d'information.

Méthodes par apprentissage statistique. Les méthodes par apprentissage statistique tentent d'offrir une solution aux problèmes de constitution des règles. Dans ce type d'approche, l'utilisateur humain délègue à la machine le processus de construction des règles, en fournissant des données annotées et des exemples de sorties attendues. Dans ce type d'approche, s'il est rapide d'appliquer un modèle pour anonymiser de gros volumes de données, le temps d'annotation du corpus fourni comme base d'apprentissage ne doit cependant pas être négligé. D'autre part, pour que le modèle construit par apprentissage soit le plus robuste possible, il est nécessaire de fournir une base d'apprentissage la plus complète possible en termes de contextes d'apparition des informations à anonymiser. Sur le challenge i2b2 2006, parmi les différents formalismes utilisés, les arbres de décision et les CRF ont permis l'obtention de meilleurs résultats que les SVM. Toutefois, l'utilisation de ressources externes complémentaires d'ordres sémantique et syntaxique ont permis aux équipes qui en ont fait usage d'obtenir de meilleurs résultats. Si les expériences de biais du rappel ont permis d'augmenter la couverture des informations devant être anonymisées, le corpus final a perdu en qualité, avec des conséquences limitées sur les traitements ultérieurs appliqués sur le corpus ainsi anonymisé.

Méthodes hybrides. Enfin, l'hybridation de plusieurs méthodes, parce qu'elle combine des méthodes aux apports distincts, permet d'obtenir des résultats de meilleure qualité. L'hybridation existe sous différents aspects, tous donnant de bons résultats : l'utilisation des méthodes symboliques pour produire les attributs utilisés par l'apprentissage pour enrichir la base d'apprentissage, l'utilisation des méthodes symboliques comme pré- et post-traitements aux méthodes par apprentissage pour corriger les erreurs statistiques, l'utilisation de plusieurs systèmes en cascade, soit en succession l'un après l'autre, soit en parallèle, avec une procédure de vote final pour décider s'il importe de conserver l'anonymisation réalisée par chaque système. L'hybridation des systèmes constitue actuellement le champ de recherche le plus actif et permet, en matière d'anonymisation automatique, d'obtenir des résultats de qualité.

Chapitre 3

L'évaluation

« Propre, soigneux, travailleur, silencieux, digne de confiance... » Grand Dieu ! Quel genre de monstre veulent-ils donc ? Je crois que jamais je ne pourrais travailler pour une firme dotée d'une telle vision du monde.

La conjuration des imbéciles
JOHN KENNEDY TOOLE

Sommaire

3.1 Introduction

Dans ce chapitre, nous introduisons les différentes mesures d'évaluation existantes et utilisées dans les tâches d'anonymisation automatique.

Il existe deux types d'évaluation, l'évaluation humaine et l'évaluation automatique. Le choix d'un type d'évaluation est dépendant des traitements appliqués au corpus et du résultat que l'on souhaite évaluer. Par exemple, évaluer la lisibilité d'un corpus anonymisé automatiquement ne semble possible que de manière humaine, éventuellement avec une aide semi-automatique.

Dans le cas d'une évaluation humaine, le processus d'évaluation sera long et coûteux. Dans le cas où plusieurs humains jugeraient de la qualité du corpus, les avis pourront diverger sur la qualité obtenue, conduisant à des étapes ultérieures de confrontation des résultats, de discussion et d'adjudication. Cependant, dans le contexte d'une anonymisation d'un corpus médical avec pour objet la sortie de ce corpus hors de l'hôpital, une étape de vérification humaine des anonymisations produites sur ce corpus se révèle particulièrement indispensable.

Les mesures d'évaluation automatique que nous présentons dans ce chapitre ont été créées pour évaluer les résultats de classification automatique. Elles sont cependant utilisées pour évaluer le résultat d'une anonymisation automatique, tâche qui est alors conçue comme une tâche de classification des informations à anonymiser dans des catégories prédéfinies *(nom, prénom, date, etc.)*. Il existe des mesures d'évaluation mono-classe, pour évaluer une seule catégorie *(rappel, précision, spécificité, etc.)*, et multi-classes, pour évaluer plusieurs catégories en même temps *(macro et micro-mesures)*. On oppose également les mesures non pondérées *(les mesures précédentes)* aux mesures pondérées *(F-mesure, indice de Jaccard, etc.)*.

Alors que ces mesures ne permettent d'évaluer que la catégorisation de l'information traitée, d'autres mesures prennent en compte à la fois la catégorisation et les frontières de l'information. Compte tenu de la priorité que l'on accorde, soit à la catégorisation, soit aux frontières, il est possible d'effectuer une évaluation avec un relâchement de contraintes. Ainsi, en matière d'anonymisation, l'évaluation peut être assouplie si la portion anonymisée par le système intègre pleinement la portion annotée dans la référence.

Dans le cas d'une évaluation automatique, il faut pouvoir bénéficier d'une référence à laquelle comparer le corpus anonymisé. Le corpus de référence, appelé « *gold standard* » en anglais, correspond à l'état du corpus dans lequel on souhaiterait le trouver au terme d'une phase d'anonymisation parfaite. La constitution de cette référence est une tâche également longue et coûteuse, avec là aussi des phases d'adjudication nécessaire si plusieurs humains ont produit la référence, ce qui est préférable pour garantir la qualité du corpus de référence.

Afin d'évaluer la qualité du travail d'annotation réalisé par plusieurs humains, il existe différents coefficients d'accord inter-annotateurs appartenant à la famille des kappa. Ces coefficients permettent de rapporter les accords observés entre deux annotateurs aux accords que l'on obtiendrait par un simple tirage au hasard. Nous présentons ces différents coefficients au regard de la tâche d'anonymisation automatique conduite dans ce travail.

3.2 Les mesures d'évaluation

Quelle que soit la tâche considérée, une évaluation des résultats produits s'avère nécessaire. Comme le rappellent [Manning et Schütze, 2000], l'utilisation de mesures d'évaluation a permis de démontrer les améliorations des performances de systèmes, notamment dans le cas de tâches applicatives. Les auteurs indiquent par ailleurs que des évaluations répétées d'un système permettent de créditer l'amélioration globale d'un système à l'un de ses composants, en cas d'augmentation des résultats.

3.2.1 Introduction

Afin de mesurer les performances d'un système par rapport à une référence, il est nécessaire, pour chaque catégorie d'information, d'établir une matrice de confusion. Cette matrice, présentée sous la forme d'un tableau à deux dimensions, consiste à relever les accords et les désaccords de catégorisation entre l'hypothèse et la référence. Particulièrement utilisée dans les domaines de la biologie et de la médecine, elle permet de faciliter les études épidémiologiques. [1] Une matrice met en évidence quatre décomptes, à partir desquels il est possible de calculer des mesures complémentaires que nous détaillerons dans la suite de ce chapitre.

Sur une tâche d'étiquetage de documents textuels *(étiquetage des tokens en parties du discours, étiquetage en entités nommées, etc.)*, les quatre décomptes se définissent comme dans la matrice suivante (tableau 3.1) et la figure 3.1.

		RÉFÉRENCE	
		Étiqueté	Non étiqueté
HYPOTHÈSE	Étiqueté	Vrais positifs	Faux positifs
	Non étiqueté	Faux négatifs	Vrais négatifs

TABLE 3.1 – Matrice de confusion adaptée au Traitement Automatique des Langues

- Le nombre de « vrais positifs » correspond, pour une étiquette donnée, au nombre d'éléments étiquetés de la même manière dans l'hypothèse et la référence ;
- Le nombre de « faux positifs » (également appelés « faux succès » ou « erreurs de type I ») correspond, pour une étiquette donnée, au nombre d'éléments étiquetés dans l'hypothèse qui sont absents de la référence (pas d'étiquetage ou étiquette différente) ;
- Le nombre de « faux négatifs » (également appelés « faux rejets » ou « erreurs de type II ») se rapporte, pour une étiquette donnée, au nombre d'éléments étiquetés dans la référence qui sont absents de l'hypothèse (pas d'étiquetage ou étiquette différente) ;
- Enfin, le nombre de « vrais négatifs » correspond à l'ensemble des éléments qui sont absents de l'hypothèse et de la référence.

1. http://consultation.demotis.org/glossaire/épidémiologie : « *L'épidémiologie est l'étude des facteurs influant sur la santé et les maladies des populations humaines. Il s'agit d'une science qui se rapporte à la répartition, à la fréquence et à la gravité des états pathologiques.* »

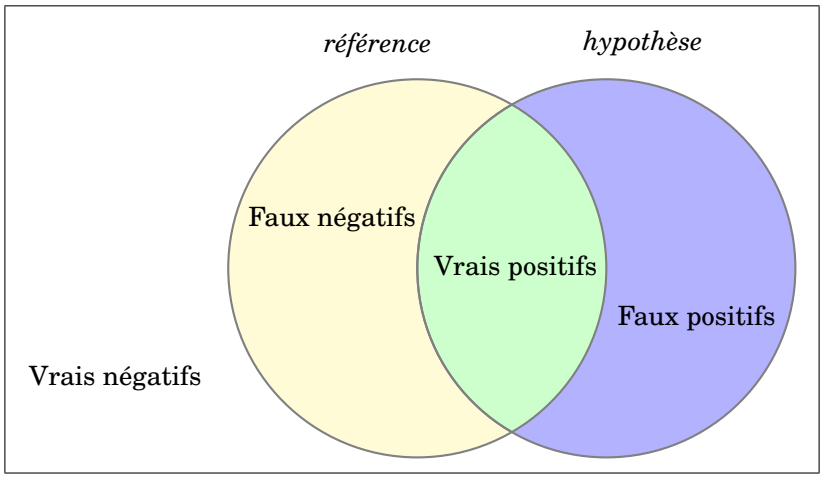


FIGURE 3.1 – Représentation des types de réponses selon la théorie des ensembles

3.2.2 Évaluation mono-classe

Les performances des systèmes issus du traitement automatique des langues sont généralement évaluées au moyen de deux mesures complémentaires, le rappel et la précision, et d'une troisième mesure qui combine les deux précédentes, la F-mesure. Une autre mesure est également utilisée, notamment en biologie : l'exactitude.

Parce que ces mesures ont été appliquées à l'origine pour évaluer des systèmes de recherche d'information, les mesures qui reposent sur les vrais négatifs (dont le nombre est forcément élevé) sont assez peu utilisées. Une description complète des différentes mesures d'évaluation présentées infra est disponible dans l'ouvrage de [Manning et Schütze, 2000].

Mesures non pondérées

Rappel. Le « rappel » *(recall)*, également appelé « sensibilité » *(sensitivity)* ou « taux de vrais positifs », est une mesure quantitative (formule 3.1). Elle mesure le nombre d'éléments correctement étiquetés par le système (vrais positifs) rapporté au nombre d'éléments étiquetés dans la référence (vrais positifs et faux négatifs).

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs + faux négatifs}} \tag{3.1}$$

La mesure complémentaire du rappel est le « silence » ; il correspond au nombre de faux négatifs. Il s'agit des éléments d'une catégorie présents dans la référence qui n'auront pas été étiquetés par le système

Précision. La « précision » *(precision)*, également appelée « valeur prédictive positive », est une mesure qualitative (formule 3.2). Elle mesure le nombre d'éléments correctement étiquetés par le système (vrais positifs) rapporté au nombre total d'éléments étiquetés par le système (vrais et faux positifs).

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{vrais positifs + faux positifs}} \tag{3.2}$$

La mesure complémentaire de la précision est le « bruit » ; il correspond au nombre de faux positifs. Il s'agit des éléments étiquetés par le système qui n'auraient pas dû l'être.

Indice de Jaccard. Cet indice est une mesure statistique proposée par Paul Jaccard (1868–1944) pour comparer la similitude et la diversité entre deux populations *(également appelé « indice de Tanimoto »)*. Il correspond au nombre de vrais positifs rapporté à la somme des vrais positifs, faux positifs et faux négatifs (formule 3.3). En complément, la distance de Jaccard mesure la diversité entre deux ensembles : Distance $= 1 -$ indice de Jaccard

$$\text{Jaccard} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs} + \text{faux négatifs}} \tag{3.3}$$

Exactitude. L'« exactitude » *(accuracy)*, que l'on retrouve également sous le nom d'« indice de Sokal et Michener » parmi la famille des indices de similitude (formule 3.4) est souvent utilisée en complément de la F-mesure, y compris pour évaluer les résultats des systèmes issus du traitement automatique des langues. Elle correspond au nombre de prédictions justes rapporté au nombre total de prédictions.

$$\text{Exactitude} = \frac{\text{vrais positifs + vrais négatifs}}{\text{vrais positifs + vrais négatifs + faux positifs + faux négatifs}} \tag{3.4}$$

Cette mesure se révèle pourtant inadaptée du fait du nombre élevé de vrais négatifs (c.-à-d., des éléments qui ne sont annotés, ni dans la référence, ni dans l'hypothèse). Autrement dit, cette mesure ne permet pas la comparaison de faibles valeurs de vrais positifs, faux positifs et faux négatifs parmi un taux élevé de vrais négatifs. Ainsi, en matière de recherche d'information, il est aisé d'estimer qu'une très large majorité de documents (proche de 99,9 %) ne correspond pas à la requête de l'utilisateur. Calculer l'exactitude ne permet pas d'accorder du sens aux résultats obtenus, ce qui demanderait qu'on accorde davantage d'intérêt aux éléments qui auront été ramenés.

Spécificité. La « spécificité » *(specificity)*, également appelée « sélectivité » *(selectivity)* ou « taux de vrais négatifs » (formule 3.5) permet d'apprécier le nombre d'éléments non annotés dans un corpus. Cette mesure reste cependant principalement utilisée par la biologie et la médecine plutôt que par le traitement automatique des langues, pour les mêmes raisons de taux élevé de vrais négatifs parmi les données traitées.

$$\text{Spécificité} = \frac{\text{vrais négatifs}}{\text{vrais négatifs + faux positifs}} \tag{3.5}$$

Mesures pondérées

F-mesure. Afin de faciliter la comparaison directe des performances de plusieurs systèmes, notamment dans le cadre de campagnes d'évaluation en classification automatique, a été créée une mesure qui combine le rappel et la précision. La « F-mesure » *(F-measure)* ou « F-score » (voire « mesure-F » dans la littérature francophone au Québec), est la moyenne harmonique pondérée du rappel et de la précision (formule 3.6). La valeur accordée à β permet, soit d'équilibrer le rappel et la précision (avec $\beta = 1$, on accorde un poids équivalent aux deux mesures), soit d'accorder plus d'importance à l'une des deux mesures (on pondère une mesure par rapport à l'autre) : le rappel au détriment de la précision (si $\beta > 1$) ou inversement, la précision au détriment du rappel (si $\beta < 1$).

$$\text{F-mesure} = \frac{(1 + \beta^2) \times \text{ précision } \times \text{ rappel}}{\beta^2 \times \text{ précision } + \text{ rappel}} \tag{3.6}$$

Ainsi, pour une tâche d'anonymisation, on privilégiera un système qui aura totalement anonymisé les éléments devant l'être (objectif prioritaire d'une telle tâche), même si des éléments auront été anonymisés alors qu'ils n'auraient pas dû l'être (on parle alors de « sur-anonymisation »). Autrement dit, on tolèrera du « bruit » (des éléments anonymisés en excès) dès lors que toutes les informations devant être anonymisées l'auront été. Dans ce cas, on accordera plus d'importance au rappel (par exemple, avec $\beta = 2$, on parle alors de F_2-mesure par opposition à la F_1-mesure lorsque $\beta = 1$).

A contrario, pour une tâche de recherche d'information (un moteur de recherche par exemple), on privilégiera un système qui fournit, parmi les premiers résultats qu'il renvoie, les documents correspondant le mieux à la requête de l'utilisateur. Peu importe que le système fournisse tous les documents qui correspondent à la requête. Autrement dit, un utilisateur préférera disposer de documents valides parmi les premiers résultats retournés plutôt que de disposer de tous les documents, même s'ils correspondent partiellement à sa requête. Dans ce cas, on privilégiera la précision sur le rappel (par exemple, avec $\beta = 0,5$).

S'il est possible de pondérer le rappel ou la précision en faisant varier la valeur du paramètre β, il peut être intéressant de prendre en compte plus précisément deux taux : (i) le nombre de sur-annotation *(les faux positifs, le « bruit »)*, et (ii) le nombre de sous-annotation *(les faux négatifs, le « silence »)*.

Indices de similitude. Un ensemble d'indices de similitude a été défini pour quantifier le degré d'association de deux éléments. Tous ces indices renvoient des résultats compris dans un intervalle [0..1]. Nous ne renseignons que les indices qui ne reposent pas sur les vrais négatifs, cette catégorie d'éléments n'ayant aucun sens pour la problématique qui nous intéresse.

Dans le contexte d'une assignation de codes ICD-9, [Pestian et al., 2007] ont utilisé l'indice de Jaccard pour pénaliser les taux de sur- et sous-annotation. Cette pénalisation est justifiée par les répercussions économiques et juridiques qui touchent : (i) à l'absence de codage *(chaque code absent engendre un manque à gagner pour l'hôpital)* ou (ii) à l'excès de codage *(chaque code en excès coûte trois fois le gain d'un codage correct et entraîne un risque de poursuite potentielle pour fraude)*.[2]

L'indice de Dice est une mesure de similitude dérivée de l'indice de Jaccard. Elle accorde un poids deux fois plus élevé aux éléments partagés par l'hypothèse et la référence *(les vrais positifs)* et est définie comme suit (formule 3.7). Cette formule se révèle identique à celle de la F-mesure lorsque la valeur attribuée au paramètre β de la F-mesure est de 1.

$$\text{Dice} = \frac{2 \times \text{vrais positifs}}{2 \times \text{vrais positifs} + \text{faux positifs} + \text{faux négatifs}} \tag{3.7}$$

Alors que l'indice de Dice privilégie les accords en leur accordant un poids plus important, l'indice de Sokal et Sneath privilégie au contraire les désaccords en pondérant les éléments spécifiques à l'hypothèse *(les faux positifs)* et à la référence *(les faux négatifs)* (formule 3.8).

$$\text{ISS} = \frac{\text{vrais positifs}}{\text{vrais positifs} + 2 \times \text{faux positifs} + 2 \times \text{faux négatifs}} \tag{3.8}$$

2. « *The penalty for under-coding is simple–the hospital loses the amount of revenue that it would have earned if it had assigned the code. The regulations under which coding is done enforce an automatic over-coding penalty of three times what is earned from the erroneous code, with the additional risk of possible prosecution for fraud.* »

3.2.3 Évaluation multi-classes

En matière de repérage d'entités nommées — ou d'anonymisation (qui peut être vue comme l'application d'un REN) —, nous sommes confrontés au fait que les systèmes vont traiter plusieurs classes. Ces classes correspondent aux différentes étiquettes qui seront apposées sur les tokens du corpus : soit les différents types d'entités nommées *(personnes, fonctions, organisations, lieux, etc.)*, soit les différentes catégories d'un système d'anonymisation *(adresse, nom, prénom, téléphone, ville, etc.)*.

Les mesures précédemment définies s'appliquent sur des classes uniques. Dans le cas de systèmes multi-classes, il est nécessaire de procéder à une évaluation en deux étapes. Dans un premier temps, les mesures retenues sont calculées pour chaque classe (par exemple, calcul du rappel et de la précision sur la classe *nom*, puis sur la classe *prénom*, etc.). Dans un second temps, une moyenne sera calculée sur la base de l'ensemble des mesures calculées à l'étape précédente. On distingue deux types de moyennes : la macro-moyenne et la micro-moyenne.

Macro-moyenne

La macro-moyenne consiste à d'abord calculer les valeurs de rappel et de précision sur chacune des n classes avant d'en effectuer une moyenne (formules 3.9 et 3.10). Dans ce mode de calcul, puisque les mesures sont calculées sur chaque classe, la macro-moyenne accorde un poids égal à chaque classe. Les chiffres obtenus permettent d'accorder plus de sens à la qualité de la classification réalisée. Cette mesure n'est donc pas sensible aux différences d'effectifs par classe.

$$\text{Macro-rappel} = \frac{\sum_{i=1}^{n} \left(\frac{\text{vrais positifs}(i)}{\text{vrais positifs}(i)+\text{faux négatifs}(i)} \right)}{n} \tag{3.9}$$

$$\text{Macro-précision} = \frac{\sum_{i=1}^{n} \left(\frac{\text{vrais positifs}(i)}{\text{vrais positifs}(i)+\text{faux positifs}(i)} \right)}{n} \tag{3.10}$$

Micro-moyenne

À l'inverse, la micro-moyenne consiste à faire la somme des vrais positifs, faux positifs et faux négatifs avant de calculer les valeurs de rappel et de précision (formules 3.11 et 3.12). Dans ce mode de calcul, on attribue un poids équivalent à chaque élément mesuré, indépendamment de sa classe d'appartenance. Ce type d'égalité entre éléments a pour conséquence de privilégier les classes composées d'un nombre élevé d'individus au détriment des classes faiblement représentées. Une classe composée d'un effectif important d'individus comptera donc davantage dans la micro-moyenne. Ainsi, en matière d'évaluation, un système évalué en termes de micro-moyenne sera particulièrement dépendant de sa performance sur les classes à fort effectif.

$$\text{Micro-rappel} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux négatifs}(i)} \tag{3.11}$$

$$\text{Micro-précision} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux positifs}(i)} \tag{3.12}$$

3.2.4 Évaluation conjointe des typage et frontières

Slot Error Rate

Puisque l'ensemble des mesures présentées précédemment ne prend en compte que la réussite et l'erreur de classification (d'un document dans une classe, d'un token dans une catégorie, etc.) et pas les indications de frontières, il est nécessaire de recourir à une nouvelle mesure qui tient compte de ces deux aspects. Cet objectif a prévalu dans la mise en place du « Slot Error Rate » (formule 3.13), une mesure composite qui tient compte à la fois des erreurs de typage et des erreurs de frontières [Makhoul et al., 1999]. Cette mesure repose sur une énumération des erreurs. Elle consiste à faire la somme des différents types d'erreurs rencontrées dans l'hypothèse, rapportée au nombre d'éléments annotés dans la référence (des « slots »).

$$\text{Slot Error Rate} = \frac{\mathrm{D} + \mathrm{I} + \mathrm{TF} + 0,5 \times (\mathrm{T} + \mathrm{F})}{\mathrm{R}} \tag{3.13}$$

Le calcul du Slot Error Rate repose sur les éléments suivants :

- D, le nombre de délétions, c.-à-d. le nombre de réponses attendues dans la référence mais non ramenées par l'hypothèse (faux négatifs) ;
- I, le nombre d'insertions, c.-à-d. le nombre de réponses ramenées par l'hypothèse mais non attendues dans la référence (faux positifs) ;
- T, le nombre d'erreurs de typage seules, c.-à-d. le nombre de réponses typées incorrectement mais avec de bonnes frontières ;
- F, le nombre d'erreurs de frontières seules, c.-à-d. le nombre de réponses avec erreur de frontière mais typées correctement ;
- TF, le nombre d'erreurs combinant typage et frontières, c.-à-d. le nombre de réponses attendues avec type et frontières incorrectes ;
- et R, le nombre de réponses attendues par la référence (vrais positifs + faux négatifs).

La valeur chiffrée dans la formule est une pénalité qui permet d'être plus ou moins strict vis à vis des erreurs de frontières et de typage. Précisons toutefois que dans le cadre des campagnes d'évaluation en entités nommées étendues du programme Quaero, une pénalité de 0,5 a été choisie pour T et F.

Avec une pénalité fixée à 0,5, les erreurs de type seul (T) ou de frontière seule (F) comptent pour un demi-point, tandis qu'une erreur conjointe de type et de frontière (TF) comptera pour un point. Avec ce choix, le coût des erreurs augmente de manière proportionnelle au nombre d'erreurs.

Porter la pénalité à 1 revient à considérer que les erreurs simultanées de type et de frontière n'engendrent pas une addition des coûts. Ainsi, une erreur de type seul (T) ou de frontière seule (F) coûte autant qu'une erreur combinant type et frontière (TF). Avec une pénalité fixée à 1, les erreurs simples (T ou F) coûtent cependant deux fois plus cher qu'avec une pénalité fixée à 0,5, ce qui peut conduire à une augmentation plus rapide du score final.

Un système idéal bénéficiera d'un Slot Error Rate le plus proche possible de zéro, dans le sens où, ne commettant aucune erreur (ni insertion, ni délétion, ni erreur de type, ni erreur de frontière), le numérateur dans la formule sera nul, avec pour conséquence mathématique un score final nul également.

Évaluation par relâchement de contraintes

Si la mesure du Slot Error Rate présente l'avantage de prendre en compte à la fois les erreurs de typage et de frontière, elle peut se révéler pénalisante lorsque les erreurs de frontière concernent des éléments peu significatifs tels que les mots outils. Ainsi, si l'on considère l'exemple 12a, on cherchera à anonymiser le prénom et le nom du patient. Que le déclencheur « Monsieur » soit inclus dans la portion anonymisant le prénom ne change rien au fait que l'information principale est anonymisée.

(12)

a. Monsieur [prénom Michael] [nom Stipe] ...

b. [prénom Monsieur Michael] [nom Stipe] ...

Les mesures d'évaluation ayant été définies pour évaluer les résultats d'une tâche de classification automatique, elles pénaliseront la présence des termes étrangers dans la portion annotée. Dans cette optique, [Ferrández et al., 2012] ont proposé d'évaluer les résultats en effectuant un relâchement de contraintes sur les frontières, en parlant de référence « pleinement contenue » (« *fully contained* »). Dans cette perspective, l'anonymisation sera considérée comme étant correcte dès lors que la portion anonymisée inclut l'ensemble de la portion couverte par la référence. Ainsi, les mots outils intégrés dans la portion anonymisée par le système ne comptent pas pour une mauvaise réponse, dès lors que l'ensemble des informations devant être anonymisé l'a été (exemple 12b). Dans le cadre de notre travail, nous envisageons également de recourir à ce type d'évaluation qui représente plus justement les objectifs de l'anonymisation.

3.2.5 Application à l'anonymisation

Afin d'illustrer le fonctionnement des différentes mesures précédemment exposées, nous appliquons l'évaluation sur l'extrait d'un document du corpus Akenaton (voir section 4.3.3). Nous précisons que dans cet exemple, les informations nominatives et numériques d'origine ont été modifiées pour les besoins de la présentation dans ce mémoire *(pseudonymes et antidatation)*. Les données cliniques sont en revanche correctes. La référence a été constituée en respectant les principes présentés dans le guide d'annotation « *Anonymisation de documents cliniques* » (voir annexe A) que nous avons établi pour permettre la constitution d'un corpus de référence.

Corpus

Nous donnons dans l'exemple 13 les différences d'anonymisation produites entre un système *(l'hypothèse)* et la référence.

(13) Monsieur [prénom Michael] [nom Stipe] ([date 04.01.60]) est malheureusement revenu dans le service du [date 28 avril au [date 5 mai 1993]] pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en [date juin 92].
Cet infarctus est survenu alors que le patient était resté depuis lors asymptomatique avec des épreuves d'effort de bon niveau et négatives. L'infarctus a été vu en cardiologie à la sixième heure. Le contrôle coronarographique effectué à l'entrée révélait une occlusion de la coronaire droite à la jonction des

segments I et II en un site où la coronarographie de [date juin 92] ne révélait qu'une irrégularité minime. L'artère a été rapidement désobstruée. Le pic de CPK a atteint [date 1957] le lendemain avec des CPK à 110. L'évolution a été simple. L'échocardiographie effectuée dès le [date 29 avril] ([prénom P.] [nom BUCK]) retrouvait une hypokinésie septo-apicale avec dyskinésie apicale localisée, et une akynésie postéro-inférieure, une dilatation moyenne du VG avec une altération de la fonction systolique globale : la fraction d'éjection ventriculaire gauche étant évaluée à 40 %, ce qui était retrouvé lors de la sortie d'hospitalisation en [date juillet]. Il existe un possible petit thrombus apical mural, et par ailleurs une fuite mitrale modérée. L'épreuve d'effort effectuée au sixième jour a permis de soutenir la charge de 150 W pendant 3 mn et d'atteindre la fréquence cardiaque de 133/mn (sous bêta-bloquant). Elle est restée négative cliniquement et électriquement. Il s'agit d'un très bon niveau d'effort superposable aux résultats des tests précédents. L'enregistrement Holter des 24 H ne retrouve pas d'arythmie significative. L'ECG haute amplification ne retrouve pas de potentiel tardif ventriculaire.

Par commodité de lecture, nous avons encadré les informations de boîtes colorées indiquant le type d'information traité. Le code couleur utilisé est le suivant :

- l'encadrement en [vert] représente les informations devant être anonymisées qui ont été traitées de manière identique entre la référence et l'hypothèse *(vrais positifs)*, par exemple le prénom « *Michael* », les noms « *Stipe* » et « *BUCK* », les dates « *04.01.60* » et « *juin 92* » ;
- l'encadrement en [rouge] représente les éléments présents dans l'hypothèse mais absents de la référence *(faux positifs)* ; ils correspondent aux ajouts du système, par exemple la mesure « *1957* » anonymisée comme une date ;
- l'encadrement de couleur [ocre] représente les éléments présents dans la référence qui n'ont pas été traités dans l'hypothèse *(faux négatifs)* ; il s'agit d'oublis du système, par exemple les dates « *29 avril* » et « *juillet* » ou encore l'initiale composant le prénom « *P.* » ;
- un double encadrement [ocre et [rouge]] représente les annotations partielles *(des erreurs de frontières et/ou de typage)* : la portion en rouge a été traitée par le système alors que la référence attendait une portion plus large, celle de couleur ocre. Du point de vue d'une tâche de classification, ce cas de figure combine à la fois un faux positif (la portion en rouge, « *5 mai 1993* ») et un faux négatif (la portion ocre, « *28 avril au 5 mai 1993* ») ;
- pour une catégorie donnée, les éléments du texte qui ne sont pas encadrés dans cette catégorie correspondent aux *vrais négatifs*.

Évaluation

La comparaison des annotations réalisées dans l'hypothèse par rapport à celles attendues dans la référence permet de remplir le tableau 3.2 en termes de vrais positifs, faux positifs et faux négatifs. Le nombre de vrais négatifs dans chaque catégorie a été obtenu en retranchant du nombre total de tokens de l'extrait (263 tokens), le nombre de portions qui sont encadrées dans cette catégorie. Ainsi, un vrai positif peut englober plusieurs tokens et constitue une portion *(« juin 92 »)* alors qu'un vrai négatif n'inclura qu'un seul token *(« Monsieur », « malheureusement », etc.)*. Cette approche est celle suivie pour l'évaluation des résultats des systèmes de REN dans les

campagnes Quaero [Galibert et al., 2011]. L'application des différentes formules est alors possible sur la base de ces décomptes.

		Catégorie		
		date	nom	prénom
Décomptes	Vrais positifs	3	2	1
	Faux positifs	2	0	0
	Faux négatifs	3	0	1
	Vrais négatifs	255	261	261
Mono-classe	Rappel	0,500	1,000	0,500
	Précision	0,600	1,000	1,000
	Indice de Jaccard	0,375	1,000	0,500
	Exactitude	0,981	1,000	0,996
	Spécificité	0,992	1,000	1,000
	F-mesure/Indice de Dice	0,545	1,000	0,667
	Indice de Sokal et Sneath	0,231	1,000	0,333
Multi-classes	Macro-rappel	0,667		
	Macro-précision	0,867		
	Micro-rappel	0,600		
	Micro-précision	0,750		
	Slot Error Rate	0,450		

TABLE 3.2 – Décompte des différences d'annotation entre hypothèse et référence

3.2.6 Discussion

Au-delà du choix des mesures à utiliser pour évaluer les sorties d'un système — autrement dit, définir les mesures à utiliser en fonction de la tâche visée —, deux interrogations doivent être prises en compte. En premier lieu, se pose la question de la possibilité de comparer les résultats de différents systèmes entre eux. En second lieu, le sens qu'il est possible d'accorder aux résultats calculés doit être envisagé.

Il est possible de comparer les résultats d'un seul système au regard d'une référence de deux manières : soit (i) de manière chronologique pour mesurer l'évolution d'un système dans le temps *(suite à l'ajout de règles ou à la modification de règles existantes)*, soit (ii) de manière parallèle en faisant varier les paramètres *(par exemple les caractéristiques à utiliser dans un système à base d'apprentissage statistique)* pour mesurer les gains ou les pertes du système selon les paramètres sélectionnés. Dans cette comparaison, il est possible d'accorder du sens aux calculs effectués. Autrement dit, il est possible d'expliquer l'évolution des résultats calculés par les modifications apportées au système.

Inversement, la comparaison des mesures obtenues par plusieurs systèmes se révèle bien souvent complexe, si ce n'est impossible. Plusieurs raisons expliquent ces difficultés :

– la comparaison est strictement impossible si les systèmes ont été appliqués sur des jeux de données différents, les performances d'un système étant fortement liées aux caractéristiques du corpus traité : en premier lieu, le type de document clinique étudié *(comptes rendus hospitaliers, lettre de suivi, notes d'infirmières)*, en second lieu, la source des documents *(un ou plusieurs centres hos-*

pitaliers, s'agit-il du même centre pour les différents systèmes, etc.), enfin, le format des documents *(documents rédigés en langue naturelle vs. énumération de faits et de traitements médicaux)* ;
– la comparaison est également difficile si les éléments traités dans le corpus diffèrent *(uniquement les noms et prénoms du patient vs. toutes les catégories du HIPAA)* ;
– enfin, le type de traitement appliqué aux informations à anonymiser *(encadrement des informations par des balises vs. remplacement des informations par des balises XML ou des blancs typographiques)* rend également complexe la comparaison, d'autant plus qu'il ne sera pas forcément possible d'évaluer les mêmes propriétés *(uniquement la capacité des systèmes à détecter les entités ou également leur capacité à typer les entités identifiées ?)*.

En ce qui concerne le choix des mesures à utiliser pour évaluer les sorties d'un système, la tâche d'anonymisation pouvant être envisagée comme une tâche de classification, les mesures habituellement utilisées en classification automatique *(principalement, le rappel, la précision et la F-mesure, moins fréquemment l'exactitude et le slot Error Rate)* permettent de correctement évaluer les performances des anonymiseurs.

3.3 L'évaluation humaine

À l'opposé de l'évaluation automatique, qui repose sur les précédentes mesures d'une part, et sur un corpus dit de référence d'autre part, existe l'évaluation humaine, c.-à-d. réalisée directement par un humain. Le fonctionnement de l'évaluation humaine est distinct et fait intervenir différents acteurs selon le type de corpus, le domaine traité, et les finalités de recherche envisagées.

3.3.1 L'évaluation humaine en domaine oral

En matière de corpus oraux, et notamment pour des corpus d'interactions humaines, [Reffay et Teutsch, 2007] estiment que les avis de trois types d'intervenants se révèlent nécessaires pour disposer d'une évaluation de qualité.

En premier lieu, le sujet de l'anonymisation, autrement dit, la personne enregistrée ou de qui l'on parle dans le corpus et pour laquelle des informations sont renseignées dans les documents. Pour ce type de corpus, l'acteur est la personne la plus à même de juger des informations personnelles devant être anonymisées de celles, moins identifiantes, pouvant rester en clair dans les documents.

En second lieu, l'anonymisateur détenteur du corpus. Celui-ci peut juger de la possibilité d'utiliser le corpus ainsi traité au regard des traitements qui y seront appliqués et des informations nécessaires aux futurs traitements de la recherche appliquée sur ce corpus.

En dernier lieu, un chercheur étranger à la constitution et aux traitements du corpus pourra donner un avis externe sur la lisibilité du corpus produit.

3.3.2 L'évaluation humaine en anonymisation

En repérage d'entités nommées ou en classification, le concepteur d'un système cherchera à identifier tous les éléments d'une catégorie *(par exemple, toutes les oc-*

currences des prénoms d'un document). L'évaluation permettra de mettre en évidence que le système a bien identifié toutes les occurrences de la catégorie.

En anonymisation de documents cliniques, la règle qui est préconisée, comme nous l'avons vu au premier chapitre, impose d'anonymiser tous les éléments qui permettent la réidentification du patient, sans qu'une liste de catégories à traiter ne soit donnée. En matière d'anonymisation et contrairement à du repérage d'entités nommées, on ne cherchera donc pas à instancier toutes les catégories *(dans le sens où il est effectivement possible de le faire)*, mais uniquement les éléments qui permettent la réidentification du patient.

Considérer que seuls les éléments permettant la réidentification du patient doivent être anonymisés implique que des éléments — bien que relevant de catégories traitées dans d'autres documents *(par ex., un nom de ville)* — peuvent rester en clair, au motif que ces éléments laissés en clair ne permettent pas de réidentification.

On peut donc considérer qu'il existe une forme de *tolérance* quant au fait de ne pas anonymiser certains éléments, dès lors que l'anonymat du patient est garanti sans possibilité de réidentification. À l'intérieur de la classe des noms, on tolèrera davantage que le nom d'un chirurgien soit resté en clair que celui d'un patient.

Cette appréciation étant posée, émerge alors la question de l'évaluation. Compte tenu du fait que dans un document donné, l'élément d'une catégorie peut rester en clair alors que dans d'autres documents, ce même élément devra être masqué *(pour cause de possible réidentification par combinaison avec d'autres informations)*, l'évaluation classique au moyen des mesures utilisées en classification ne se révèle pas nécessairement adaptée.

Une évaluation reposant sur le jugement humain *(les informations laissées en clair dans ce document — ou utilisées en combinaison avec les informations des autres documents du dossier — permettent-elles de réidentifier le patient ?)* avec une échelle à deux valeurs *(oui/non)* serait plus adaptée, mais également plus permissive vis à vis des oublis et erreurs du système.

Cependant, afin de faciliter l'évaluation, et parce qu'il n'est pas nécessairement évident pour un humain de juger la qualité d'une anonymisation au regard de la complexité de l'appréciation de réidentification potentielle, l'ensemble des tâches d'anonymisation évalue les performances des anonymiseurs au moyen des mesures d'évaluations habituellement utilisées en TAL et présentées dans ce chapitre. Dans le cadre de nos expérimentations, nous évaluerons également les résultats de nos systèmes au moyen des mesures « classiques ».

3.4 Les accords inter-annotateurs

Lorsque plusieurs annotateurs humains accomplissent une tâche d'annotation de corpus, ils se réfèrent à un guide d'annotation pour prendre la décision d'annoter ou non un élément du texte.

Évaluer les annotations humaines. Puisque l'annotation humaine laisse une part relativement importante à l'annotateur d'apprécier personnellement s'il doit annoter ou non un élément, des comparaisons d'annotations doivent être effectuées. Afin de quantifier ces comparaisons, de multiples taux d'accord inter-annotateurs ont été définis depuis de nombreuses années.

Le principe de base des accords inter-annotateurs consiste à mettre en évidence le taux d'accord entre deux annotations — que l'on espère indépendantes — en regard

de l'accord entre des annotations que l'on obtiendrait par un simple tirage au hasard. Autrement dit, dans quelle mesure les annotations produites par des humains, sans que ces humains ne se concertent de trop (principe d'indépendance des annotations), sont-elles proches ou éloignées l'une de l'autre.

3.4.1 Processus de calcul

Il est possible de calculer des taux d'accord inter-annotateurs de différentes manières selon les objectifs visés. Chaque mode de calcul apporte une information importante sur la manière dont le corpus a été annoté.

Accord entre annotateurs. La première finalité de ces taux consiste à calculer l'accord entre deux annotateurs ayant travaillé sur le même jeu de données. Ce mode de calcul permet de mettre en évidence dans quelle mesure les deux annotateurs sont d'accord entre eux sur les annotations produites.

Il s'agit du mode de calcul le plus couramment réalisé. D'autres modes de calculs s'avèrent cependant possibles. Ces autres modes portent, soit de manière globale sur l'ensemble du corpus annoté, soit de manière individuelle au niveau des annotations produites par un seul annotateur.

Bénéfice des corrections. Un deuxième mode de calcul au niveau global consiste à évaluer le bénéfice de corrections qui auront pu être apportées sur un corpus. On calculera ainsi l'accord, sur un même jeu de données, avant et après correction.

Cohérence d'un annotateur dans le temps. Au niveau individuel, un premier mode de calcul consiste à évaluer la cohérence d'un annotateur dans le temps. Pour ce faire, on cherchera à évaluer les annotations produites par un annotateur humain à deux moments différents de la phase d'annotation, par exemple en début et en fin de la période d'annotation. Les annotations produites par cet annotateur seront comparées à la référence correspondante. Ce calcul permet ainsi de mesurer la progression de l'annotateur dans le temps et de mettre en évidence les éventuelles divergences d'annotation qui auront pu se produire.

Bénéfice de la production d'un annotateur. Puisque les annotations produites par plusieurs annotateurs doivent faire l'objet de phases d'adjudication, il est possible de mesurer le bénéfice des annotations produites par un annotateur. On mesurera ainsi l'accord entre la version annotée par cet annotateur et le résultat de la phase d'adjudication. Ce mode de calcul permet également de mettre en évidence le comportement individuel d'un annotateur par rapport à une équipe complète d'annotateurs.

3.4.2 La famille des Kappa

Présentation générale

La famille des coefficients Kappa regroupe plusieurs coefficients d'accord inter-annotateurs définis dans les années 1950 et 1960 pour évaluer les annotations produites dans le cadre de tâches de classification telles que l'étiquetage en parties du

discours [Artstein et Poesio, 2008]. Ces coefficients ont la particularité de tous partager une formule générique (formule 3.14), d'où l'appellation de « famille de coefficients ». Cette famille comprend trois principaux coefficients : S, π et κ, définis pour comparer des annotations entre deux codeurs. Une généralisation de ces coefficients a également été réalisée pour évaluer les annotations entre plusieurs codeurs. Les résultats fournis par ces coefficients varient principalement entre 0 et 1, mais il est également possible d'obtenir des résultats négatifs.

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e} \tag{3.14}$$

Les éléments constitutifs de cette formule sont les suivants :
- A_o (*observed agreement*) correspond au taux d'accords observés, c.-à-d. le pourcentage de réponses communes aux deux annotateurs ;
- A_e (*expected agreement*) correspond au taux d'accords attendus par le biais d'un tirage au hasard.

Dans cette formule, $A_o - A_e$ représente l'écart entre l'accord observé et l'écart attendu par rapport au hasard tandis que $1 - A_e$ permet de mesurer à quel point l'accord dû au hasard est atteignable. La seule différence entre les scores S, π, κ réside dans la manière de calculer l'accord attendu A_e.

Tous ces scores ne sont cependant pas forcément adaptés aux spécificités des tâches du traitement automatique des langues. Nous résumons dans les paragraphes suivants les idées maîtresses de ces différents coefficients, pour une comparaison binaire, telles qu'elles sont présentées par [Artstein et Poesio, 2008]. Nous reprenons également la terminologie employée par ces derniers pour désigner les différents coefficients.

L'accord entre deux annotateurs

Le coefficient S. Selon [Bennett et al., 1954], auteurs du coefficient S, une annotation due au hasard doit engendrer une distribution uniforme entre catégories. Autrement dit, le même nombre d'éléments doit figurer dans chaque catégorie. Dans cette conception, le calcul de l'accord aléatoire repose sur le nombre total de catégories (formule 3.15) avec n le nombre de catégories.

$$A_e^S = n \times \left(\frac{1}{n}\right)^2 \tag{3.15}$$

Ce coefficient apparaît difficilement utilisable pour évaluer des tâches relevant du traitement automatique des langues. En effet, le modèle d'une distribution uniforme entre catégories n'est pas plausible dans les tâches d'étiquetage. Ainsi, sur une tâche d'anonymisation automatique, toutes les catégories ne se composent pas du même nombre d'entités (aucun document clinique ou corpus ne comprend le même nombre de mots issus de chaque catégorie : *noms, prénoms, ville, dates, code postal, etc.*).

[Artstein et Poesio, 2008] indiquent par ailleurs que, du fait de la distribution uniforme entre catégories, la valeur finale peut être artificiellement gonflée par l'ajout de catégories que les annotateurs n'utiliseront jamais (voir exemple).

Le coefficient π. De manière différente, [Scott, 1955] propose le coefficient π (également connu sous le nom de « Kappa de Carletta » [Carletta, 1996]), et considère

qu'une annotation due au hasard implique une distribution identique entre annotateurs mais différente entre catégories. Autrement dit, on doit retrouver le même nombre d'éléments par annotateur. Selon cette conception, le hasard distingue les catégories (puisque le nombre d'éléments peut varier par catégorie) mais pas les annotateurs (formule 3.16). Cette conception ne paraît pas recevable. Dans les expériences d'annotation humaine de corpus que nous avons réalisées, aucun annotateur humain n'annotait le même nombre d'entités.

Soit N le nombre de jugements pour une catégorie, la probabilité d'annotation pour une catégorie k_a sera notée $\hat{P}(k)$.

- Soit pour un annotateur, $\frac{n_{k_a}}{N}$
- Et pour deux annotateurs, $(\frac{n_{k_a}}{N})^2$

$$A_e^{\pi} = \sum_k \hat{P}(k)^2 = \sum_k \left(\frac{n_k}{N}\right)^2 \tag{3.16}$$

Le coefficient κ. Enfin, [Cohen, 1960] considère que les annotateurs interprètent différemment les instructions fournies. Cette différence d'interprétation induit un biais que le coefficient κ cherche à modéliser (formule 3.17). Cette conception se révèle plus juste et plus proche des observations faites en termes d'annotation humaine de corpus.

Soit i le nombre total d'éléments annotés, la probabilité d'annotation d'une catégorie k_a sera :

- pour un annotateur A_x : $\frac{nA_xk_a}{i}$
- pour deux annotateurs A_1 et A_2 : $\frac{nA_1k_a}{i} \times \frac{nA_2k_a}{i}$

$$A_e^{\kappa} = \sum_k \frac{nA_1k}{i} \times \frac{nA_2k}{i} \tag{3.17}$$

[Artstein et Poesio, 2008] relèvent ainsi que le coefficient π reflète les annotations attendues par des annotateurs arbitraires alors que le coefficient κ reflète ce que les annotateurs ont réellement produit.

Devant l'impossibilité de dénombrer les vrais négatifs, en particulier lorsque ce nombre est élevé, [Hripcsak et Rothschild, 2005] ont démontré qu'il était possible d'utiliser la F-mesure pour calculer des accords inter-annotateurs. La valeur ainsi obtenue tend à se rapprocher de la valeur qu'on obtiendrait avec un κ, notamment lorsque le nombre de vrais négatifs est élevé.

L'accord au-delà de deux annotateurs

Afin d'évaluer les annotations produites par plusieurs annotateurs, une généralisation des coefficients existants a été réalisée. Cette généralisation permet uniquement d'évaluer les annotations réalisées par plus de deux annotateurs. Les critiques formulées quant à l'inadéquation de certains coefficients en matière d'annotation ne disparaissent pas sous l'effet de la généralisation. Dans le cas de plusieurs annotateurs, il devient impossible de mesurer l'accord observé A_o de la même manière que pour les coefficients entre deux annotateurs, dans le sens où l'obtention d'une majorité d'accord entre tous les annotateurs se trouve fortement réduite.

Généralisation du coefficient π. Une généralisation du coefficient π de Scott a été proposée par [Fleiss, 1971] pour évaluer l'accord entre plusieurs annotateurs. Ce coefficient — nommé κ de Fleiss — a été renommé multi-π par Artstein et Poesio pour éviter toute confusion avec le κ de Cohen. Cette généralisation repose sur un accord par paires. Le taux d'accord observé A_o sur une catégorie correspond au ratio entre les paires d'accord et le nombre total de paires sur cette catégorie. Le taux d'accord attendu A_e repose également sur un accord par paires (formule 3.18). À l'image du coefficient π duquel il est dérivé, le coefficient multi-π repose sur un accord attendu qui suppose une distribution identique d'éléments entre annotateurs.

$$A_e^{\pi} = \sum_{k \in K} \left(\frac{1}{ic} n_k \right)^2 \tag{3.18}$$

Avec :
- K l'ensemble des catégories disponibles ;
- n_k le nombre d'éléments attribués à la catégorie k ;
- ic le nombre d'éléments i multiplié par le nombre d'annotateurs c.

Généralisation du coefficient κ. Une généralisation du coefficient κ de Cohen a également été réalisée — nommée multi-κ par Artstein et Poesio —, proposée par [Davies et Fleiss, 1982]. Cette généralisation repose sur la moyenne des accords entre codeurs pris deux à deux. L'accord attendu A_e^{κ} pour plusieurs annotateurs est la moyenne des valeurs A_e^{κ} sur toutes les paires d'annotateurs.

Grille de lecture

Afin d'interpréter les résultats calculés par un coefficient, plusieurs grilles de lecture ont été proposées. Nous mentionnons les trois échelles rapportées par [Artstein et Poesio, 2008].

La première grille de lecture a été proposée par [Landis et Koch, 1977]. Il s'agit d'une échelle à six valeurs avec des intervalles de 0,2 points. Selon cette échelle, l'accord sera qualifié d'*excellent* pour une valeur comprise entre 0,81 et 1, de *bon* entre 0,61 et 0,8, de *modéré* entre 0,41 et 0,6, de *médiocre* entre 0,21 et 0,4, de *mauvais* entre 0 et 0,2 et de *très mauvais* en cas de valeur négative.

Quelques années plus tard, [Krippendorff, 1980] a proposé une grille composée de trois valeurs. Plutôt que de qualifier les accords, l'auteur décrit quelle est l'utilisation qui peut être faite des annotations humaines ainsi produites. L'auteur considère alors qu'il y a *cohérence* des annotations pour une valeur comprise entre 0,81 et 1, ce premier intervalle étant commun à l'échelle de [Landis et Koch, 1977]. Pour une valeur strictement inférieure à 0,67, l'auteur considère qu'il y a *incohérence* entre les annotateurs. Enfin, un troisième intervalle, compris entre 0,67 et 0,8 implique qu'*aucune décision* ne peut être prise.

Plus récemment, [Green, 1997] a proposé une autre grille d'évaluation à trois valeurs également. L'accord est ainsi qualifié d'*élevé* pour une valeur comprise entre 0,75 et 1, de *moyen/bon* entre 0,4 et 0,75, et de *faible* entre 0 et 0,4.

Il existe bien d'autres grilles de lecture. Cependant, les trois présentées dans cette section nous permettent déjà de dresser le constat suivant. Aucune des grille ne comporte le même nombre d'intervalles et aucune ne repose sur les mêmes frontières d'intervalles. Il semble particulièrement difficile de vouloir associer du sens aux résultats chiffrés d'un accord inter-annotateurs. Pour cette raison, [Artstein et Poesio, 2008]

considèrent que si un seuil doit être retenu, alors les accords dont la valeur est supérieure à 0,8 constitueront un bon accord. En deçà de cette valeur, les annotations humaines ne pourront être considérées comme pertinentes, et un nouveau travail de ré-annotation devra être effectué.

3.5 L'interprétation des résultats

S'il est établi qu'une comparaison entre systèmes est possible, mais également entre différentes méthodes d'un même système, il est alors envisageable de mobiliser des méthodes statistiques pour interpréter les résultats. On cherchera alors à calculer la significativité statistique des résultats obtenus par deux systèmes ou deux approches différentes.

La significativité statistique permet de mettre en évidence dans quelle mesure les différences de résultats sont significatives sur le plan statistique. Le test de Student permet de calculer cette significativité. On estime généralement que les différences sont significatives si les résultats produits par ce test sont inférieurs au seuil observé de 0,05 ; autrement dit, si le résultat observé a moins de 5 % de chances d'être dû au hasard, soit une confiance de 95 % dans les résultats produits.

Un intervalle de confiance consiste à remplacer la valeur d'une estimation donnée par un intervalle dont les extrémités dépendent d'un seuil de confiance fixé a priori *(généralement fixé à 95 %)* [Dress, 2004]. Appliqué à la F-mesure, l'intervalle de confiance dans lequel évolue la F-mesure d'un système permet de représenter plus précisément les performances de ce système. Il est couramment admis que plus l'intervalle de confiance calculé est réduit, plus les performances du système pourront être qualifiées de robustes. Il est possible de calculer un intervalle de confiance de deux manières, en utilisant le test de Student, ou par les méthodes de Monte Carlo.

Dans les expériences que nous avons menées, nous avons utilisé la simulation de Monte Carlo pour calculer les intervalles de confiance dans lesquels évolue la F-mesure.

3.5.1 Le test de Student

Le premier moyen de calculer un intervalle de confiance repose sur le test de Student. L'intervalle de confiance I_c d'un échantillon (formule 3.19) est calculé autour de la moyenne $\bar{x}$ de cet échantillon, sur la base du t critique de Student t_α, de l'écart-type σ, et de la racine carrée de la population de l'échantillon $\sqrt{n}$

$$I_c = [\bar{x} - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + t_\alpha \frac{\sigma}{\sqrt{n}}] \tag{3.19}$$

Le t critique de Student t_α est déterminé depuis la table de la Loi de Student[3] pour un nombre de degrés de liberté défini par ddl $= n - 1$ et un risque α bilatéral (on définit généralement $\alpha = 0,05$ soit une confiance de 95 %).

Rappelons que l'écart-type correspond à la racine carrée de la variance $\sigma = \sqrt{\sigma^2}$, la variance non biaisée correspond à la moyenne des carrés des écarts à la moyenne $\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ et la moyenne $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

3. Ce t critique est de 1,96 lorsque les degrés de liberté dépassent 1 000 et tendent vers l'infini.

3.5.2 Les méthodes de Monte Carlo

Présentation générale

Le deuxième moyen de calculer un intervalle de confiance repose sur la simulation de Monte Carlo [Metropolis et Ulam, 1949, Bindel et Goodman, 2006]. Les méthodes de Monte-Carlo, dont le nom fait référence aux jeux de hasard pratiqués dans la Principauté de Monaco, sont des méthodes probabilistes utilisées pour réaliser des estimations sur la base de tirages au hasard.

Les résultats d'un système étant évalués sur la base d'un corpus de référence, généralement de taille restreinte dû aux coûts de constitution d'un tel corpus, la simulation de Monte Carlo permet de simuler les résultats qu'obtiendrait ce système sur un échantillon plus important. Le fonctionnement de la simulation de Monte Carlo repose sur un nombre élevé de tirages au hasard et sur la Loi des grands nombres. Elle suppose le principe d'une distribution gaussienne centrée réduite des données. Ces méthodes permettent ainsi de calculer des données θ qui correspondent à l'espérance d'une variable X réelle telles que $\theta = E[X]$ en tirant un nombre important de dérivations de cette variable X (formule 3.20). [4] Un nombre élevé de dérivations est nécessaire pour estimer la représentation la plus probable [Manning et Schütze, 2000].

$$\theta = \lim_{x \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_n := \lim_{x \to \infty} \hat{\theta}_n \tag{3.20}$$

La représentation graphique de ces tirages correspond alors à une variable aléatoire gaussienne centrée, pour laquelle on va définir une valeur de confiance. Cette valeur revient à réduire les extrémités de la gaussienne en fonction de la confiance définie. Les méthodes de Monte Carlo reviennent ainsi à calculer l'aire sous la courbe correspondant à l'espérance de la variable X (formule 3.21), avec Z une distribution gaussienne centrée réduite. [5]

$$\mathbb{P}(Z \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt \tag{3.21}$$

Application à la F-mesure

Nous avons appliqué la simulation de Monte Carlo pour calculer l'intervalle de confiance d'une F-mesure observée en suivant le protocole suivant, avec le logiciel R, pour un risque $\alpha = 0,05$:

- pour les valeurs de vrais positifs, faux positifs et faux négatifs fournies par l'utilisateur, tirage au hasard de 10 millions de valeurs autour de ces trois valeurs au moyen de la fonction Γ (`rgamma` sous R). Cette fonction suit la loi Gamma (formule 3.22) qui permet de représenter une densité de probabilité ;

$$\Gamma(\alpha) = \int_{0}^{+\infty} t^{\alpha-1} e^{-t} dt, \forall \alpha \exists \mathbb{R}_{+}^{*} \tag{3.22}$$

- calcul des valeurs de F-mesure pour ces 10 millions de triplets <VP, FP, FN> ;
- calcul des écarts absolus entre chacune de ces 10 millions de F-mesure et la F-mesure observée sur le triplet <VP, FP, FN> fourni par l'utilisateur, écarts ordonnés par ordre croissant ;

4. http://www.infres.enst.fr/~decreuse/TP/intervalle_confiance.pdf
5. http://www.iecn.u-nancy.fr/~rmarchan/Enseignement/Masterpro/ch2-montecarlo.pdf

– calcul de l'indice de confiance à partir de la moyenne des écarts précédemment calculés pondérés par la confiance retenue ;
– production de l'intervalle de confiance sur la base du précédent indice de confiance et de la F-mesure observée.

Exemple. Nous renseignons dans le tableau 3.3 les intervalles de confiance calculés sur deux jeux de données expérimentaux. Dans la première expérience, on observe un équilibre entre nombres de vrais positifs, faux positifs et faux négatifs. Dans la mesure où une entité sur deux de la référence a correctement été traitée, et une entité sur deux parmi celles ramenées par le système est correcte, le système ayant produit ces résultats obtient une F-mesure de 0,50. Dans la seconde expérience, les performances du système se sont améliorées ; neuf entités sur dix de la référence ont correctement été traitées, et neuf entités sur dix parmi celles ramenées par le système sont correctes. Ce système obtient donc une F-mesure de 0,90.

Le calcul des intervalles de confiance autour de la F-mesure sur ces deux expériences pour un risque $\alpha = 0,05$ sur 10 millions de tirages au hasard renvoie un intervalle de confiance plus réduit pour la seconde expérience que pour la première, avec un indice de confiance de 0,0890 sur la première et de 0,0608 pour la deuxième. Ces différences d'intervalles calculés s'expliquent par la distribution des données (en termes de vrais positifs, faux positifs et faux négatifs) beaucoup plus étendue dans la première expérience que dans la seconde.

	Expérience 1	Expérience 2
Vrais positifs	90	90
Faux positifs	90	10
Faux négatifs	90	10
F-mesure observée	0,50	0,90
Indice de confiance	0,0890	0,0608
Intervalle de confiance	[0,4110;0,5890]	[0,8392;0,9608]

TABLE 3.3 – Intervalles de confiance calculés sur deux jeux de données

3.6 Synthèse

On oppose l'évaluation humaine — nécessaire pour vérifier que les données personnelles d'un corpus ont bien été anonymisées avant redistribution du corpus — à l'évaluation automatique — plus rapide et fournissant davantage de détails.

Les mesures mono-classe. En matière d'évaluation automatique, les mesures utilisées pour évaluer les résultats d'une tâche de classification automatique sont reprises à l'identique pour évaluer les tâches d'anonymisation. Les mesures prenant en compte le taux de vrais négatifs, en règle générale le taux se révèle élevé et proche des 99,9 %, sont inadaptées pour représenter et évaluer une anonymisation automatique. On privilégiera donc les mesures ne reposant pas sur ces taux *(rappel, précision)*. Deux mesures pondérées permettent de combiner le rappel et la précision *(F-mesure, indice de Jaccard)* et offrent ainsi une vision globale des performances d'un système.

Les mesures multi-classes. Appliquée à plusieurs classes, l'évaluation sera possible au moyen de l'une des deux mesures multi-classes existantes. La première mesure permet d'accorder un poids égal à chaque catégorie *(macro-mesure)* et représente plus justement la qualité des anonymisations effectuées sur chacune d'entre elles. Elle est donc utile en classification automatique pour représenter, dans le détail, la catégorisation effectuée. La deuxième mesure attribue un poids équivalent à chaque individu *(micro-mesure)*, avec pour conséquence immédiate le fait de favoriser les catégories composées d'un nombre élevé d'individus au détriment des catégories comptant peu d'individus. Cependant, sachant qu'en matière d'anonymisation, c'est l'anonymisation en tant que telle qui compte et non le typage de l'information, la micro-mesure donne un aperçu global de l'évaluation effectuée.

L'évaluation du typage et des frontières. Parce qu'elles ont été créées pour la classification automatique, les mesures précédentes évaluent uniquement la catégorie associée à un élément, en partant du principe que les frontières de cet élément sont identiques entre l'hypothèse et la référence. En matière d'anonymisation automatique, il est nécessaire d'évaluer la catégorie attribuée à un élément *(nom, prénom, etc.)* mais surtout, les frontières de cet élément. Il est essentiel de pouvoir vérifier que l'élément de l'hypothèse englobe celui de l'hypothèse. La mesure du *Slot Error Rate* permet de prendre en compte le typage et les frontières sous la forme d'une mesure composite. Cependant, toutes les erreurs de frontière ne se valent pas. Si l'hypothèse recouvre partiellement la référence, l'évaluation doit être stricte. En revanche, si l'hypothèse déborde des frontières de la référence *(par la présence du déclencheur de nom dans la portion anonymisée)*, un relâchement de contraintes peut être appliqué pour assouplir l'évaluation et amoindrir ou annuler la pénalité sur l'erreur de frontière.

Les accords inter-annotateurs. En matière d'évaluation des annotations humaines effectuées par plusieurs annotateurs, s'il existe plusieurs coefficients *(S, π et κ)*, il apparaît que seul le coefficient κ permet de prendre en compte les spécificités linguistiques *(une distribution différente des entités par catégorie)* et humaines *(une distribution différente par annotateur)* des annotations en corpus. Au-delà de deux annotateurs, la généralisation des coefficients existants repose sur un calcul des accords entre annotateurs pris deux à deux, puis en une moyenne des précédents accords calculés. Enfin, pour l'interprétation des résultats calculés, on retiendra qu'une valeur supérieure à 0,8 correspond à un bon accord, alors qu'en deçà de cette valeur, la qualité des accords n'est pas suffisante.

Significativité statistique et intervalles de confiance. Afin de calculer la significativité statistiques des résultats obtenus par deux systèmes différents, ou par deux configurations différentes d'un même systèmes, il importe de calculer l'intervalle de confiance dans lequel évolue les différentes F-mesures calculées. À cet effet, il est possible de calculer des intervalles de confiance fondés sur le test de Student, ou bien en utilisant les méthodes reposant sur un nombre élevé de tirages au hasard *(méthodes de Monte Carlo)*. Dans nos expériences, nous avons évalué les intervalles de confiance de nos différentes F-mesures en utilisant la simulation de Monte Carlo, fondée sur dix millions de tirages au hasard. Nous avons pour cela utilisé une fonction développée pour l'occasion dans le logiciel R.

Conclusion de la première partie

Dans cette première partie, nous avons passé en revue les différentes méthodes actuelles permettant de résoudre la problématique de l'anonymisation automatique de documents cliniques. À l'image des travaux en fouille de textes, ces méthodes se répartissent en deux grandes familles. Les méthodes symboliques d'une part, fondées sur l'utilisation de patrons syntaxiques et de listes, et les méthodes par apprentissage statistique d'autre part. Les travaux actuels reposent sur l'hybridation de ces deux méthodes, soit en inscrivant ces méthodes dans un processus complémentaire *(une méthode suivie d'une autre)*, soit dans un processus combinatoire *(l'enrichissement des méthodes par apprentissage au moyen des ressources et éléments produits par les méthodes symboliques)*.

❧

Nous avons ensuite détaillé les différentes mesures d'évaluation utilisées dans les travaux du traitement automatique des langues. Ces mesures ayant été conçues pour évaluer les résultats produits par des outils de recherche d'information, nous avons présenté les limites de l'application de ces mesures pour des tâches telles que le repérage d'entités nommées ou l'anonymisation automatique qui font intervenir à la fois une catégorisation *(l'affectation d'une étiquette qui type l'information traitée)* et un bornage *(la délimitation de la portion sur laquelle l'étiquette est affectée)*. Deux solutions existent pour prendre en compte ces aspects : la mesure du *Slot Error Rate* d'une part, l'évaluation englobante par relâchement de contraintes d'autre part. Nous avons également abordé les moyens d'apprécier la confiance dans les évaluations effectuées, soit par le biais de tests de significativité statistique, soit par la simulation de Monte Carlo.

❧

Dans la partie suivante de ce manuscrit, nous nous proposons de détailler les expériences que nous avons menées pour traiter de la problématique de l'anonymisation de documents du domaine médical. Les modalités d'évaluation présentées dans cette première partie seront mises en application pour mesurer les performances des systèmes utilisés.

Deuxième partie

Expérimentations

Introduction de la deuxième partie

Dans cette deuxième partie, nous avons rassemblé les différentes expériences que nous avons menées en matière d'anonymisation automatique de documents cliniques.

❧

Nous introduisons cette partie par une présentation du guide d'annotation que nous avons produit pour constituer manuellement le corpus de référence. Un guide d'annotation décrit l'objet d'étude *(que veut-on faire ?)*, présente les règles d'annotation *(quelles règles utiliser ?)* accompagnées d'exemples d'annotation *(comment annote t-on le corpus ?)*. Il précise également les objectifs visés *(pour quelle raison faire ces annotations ?)* et fournit éventuellement une description des données et des prétraitements effectués. Le corpus de référence manuellement constitué à partir de ce guide d'annotation a servi pour l'évaluation future des résultats produits lors des différentes expériences. Nous présentons par la suite les corpus sur lesquels nous avons appliqué nos méthodes et la procédure suivie pour produire le corpus de référence. Nous introduisons également brièvement les outils mobilisés pour préparer les corpus et évaluer les résultats.

❧

Nous détaillons ensuite les démarches d'anonymisation que nous avons menées reposant sur les méthodes symboliques. Nous détaillons ainsi les trois démarches que nous avons suivies : un rappel des premières approches de l'anonymisation réalisées en 2002, puis les deux expériences récentes, la première sur la tentative de francisation d'un outil existant, la seconde sur la création *ex nihilo* d'un outil dédié.

❧

Nous présentons par la suite les expériences que nous avons menées en matière d'apprentissage statistique et d'hybridation. Nous introduisons dans un premier temps le protocole expérimental que nous avons suivi, notamment en matière de validation croisée. Après avoir présenté les outils à notre disposition reposant sur le formalisme des CRF, nous exposons les paramètres de configuration que nous avons retenus pour l'outil utilisé dans nos expériences. Nous poursuivons en présentant les différentes expériences que nous avons menées.

❧

Enfin, nous terminons cette partie par un chapitre portant sur l'évaluation des résultats produits par chacune des méthodes. Nous discutons alors des avantages et des inconvénients de chaque type de méthode.

Chapitre 4

Corpus et matériau utilisés

— Ignatius, qu'est-ce que c'est que toutes ces saletés sur le plancher ?
— C'est ma vision du monde que tu vois là. Il reste à l'organiser en un tout cohérent, alors fais attention où tu mets les pieds.

La conjuration des imbéciles
JOHN KENNEDY TOOLE

Sommaire

4.1 Introduction

Dans ce chapitre, nous présentons les corpus sur lesquels nous avons élaboré puis testé nos différents systèmes d'anonymisation automatique. Dans une première section, nous présentons les guides d'annotation en général, et celui utilisé en anonymisation pour produire les corpus de référence nécessaires aux évaluations des résultats produits par les systèmes.

Dans une seconde section, nous abordons les corpus sur lesquels nous avons plus spécifiquement travaillé. Un premier corpus de comptes rendus hospitaliers et de lettres de suivi dans le domaine de la cardiologie, et un deuxième corpus en fœtopathologie. Pour chaque corpus, nous rappelons quelles sont les principales caractéristiques des documents et l'usage qui doit être fait des corpus au terme de l'anonymisation, cet usage conditionnant le maintien en clair de certaines informations.

4.2 Les guides d'annotation

4.2.1 Introduction

Dans cette section, nous présentons les caractéristiques générales d'un guide d'annotation en termes de types d'information devant figurer dans un guide et les raisons qui justifient la présence de ces éléments.

Nous poursuivons sur la présentation du guide d'annotation que nous avons défini pour l'anonymisation de corpus. Ce guide a été utilisé pour réaliser le corpus de référence du corpus de cardiologie afin d'évaluer les performances des systèmes appliqués sur ce corpus.

4.2.2 Présentation

Un guide d'annotation est un manuel dont l'objectif consiste à présenter à des humains les principes généraux et restrictions spécifiques à respecter pour annoter un corpus. Le corpus ainsi annoté servira par la suite, soit comme corpus de référence pour l'évaluation, soit comme corpus d'apprentissage pour les systèmes à base d'apprentissage statistique.

Si ce guide peut également servir comme une aide de base au développement d'un outil informatique, il importe de garder à l'esprit qu'il ne s'agit pas d'un cahier des spécifications fonctionnelles attendues. Ce manuel ne suffit pas pour développer un système, dans le sens où tous les cas de figure possibles n'y sont pas référencés et toutes les règles à implémenter n'y sont pas présentes.

La vie d'un guide d'annotation passe par plusieurs phases. Les concepteurs du guide établissent une première version composée des trois types d'éléments nécessaires dans un guide. Ce guide est alors fourni aux annotateurs humains, ces derniers annotant le corpus d'après les principes énoncés. À l'occasion de cette phase d'annotation, plusieurs retours des annotateurs seront effectués vers les concepteurs du guide, à la fois pour poser des questions *(notamment sur la manière d'interpréter une règle particulière contextuellement)* mais aussi pour faire des commentaires *(besoin d'indices supplémentaires pour savoir si une règle s'applique)*. Sur la base de ces retours, les concepteurs du guide ajusteront le guide en conséquence, donnant lieu à une nouvelle version.

Les informations obligatoires

Un guide d'annotation se doit de présenter ce qui est nécessaire pour accomplir une tâche d'annotation. [Rosset, 2010] mentionne trois types d'information qui doivent obligatoirement être présents dans le guide : (i) la description de l'objet d'étude *(permet de présenter ce que l'on souhaite annoter)*, (ii) les règles d'annotation *(présente les règles qui doivent être suivies lors de l'annotation du corpus)*, et (iii) un grand nombre d'exemples *(fournit des indices sur la manière dont le corpus doit être*

annoté). La présence d'exemples en nombre doit permettre aux annotateurs humains de s'approprier les différentes règles d'annotation, en fonction des différentes situations présentées dans ces exemples.

Les méta-informations

À ces trois catégories d'information essentielles, [Rosset, 2010] précise que des descriptions supplémentaires doivent être ajoutées : (i) la liste des principaux objectifs poursuivis dans le projet dans lequel s'inscrit cette tâche d'annotation *(permet d'expliquer pour quelles raisons les annotations sont faites)*, et (ii) une description complète des données en termes de statistiques, de pré-traitements réalisés, etc.

Les annotateurs

Le succès d'une annotation de corpus dépend de la qualité du travail accompli par les annotateurs. Cette qualité est garantie par leur bagage culturel d'origine *(formation universitaire et expérience professionnelle)*, et par la formation qu'ils auront suivie pour comprendre les principes exposés dans le guide d'annotation [Fort, 2012]. En règle générale, une partie (voire la totalité) du corpus est annotée en double par une paire d'annotateurs. Cette double annotation permet de vérifier la compréhension et l'appropriation du guide d'annotation par les annotateurs d'une part, et la cohérence des annotations entre les deux annotateurs d'autre part.

Lors de l'édition 2012 du défi i2b2 *(identification des concepts médicaux, des expressions temporelles, et des relations temporelles entre concepts et expressions)*, les corpus de documents cliniques ont été annotés en double grâce à des paires d'annotateurs. Les organisateurs de cette édition ont ainsi observé que les annotations des relations temporelles étaient de meilleure qualité lorsque la paire d'annotateurs était composée d'un clinicien et d'un linguiste, soit deux profils distincts.

4.2.3 Guide d'annotation pour l'anonymisation

Dans le cadre de ce travail de thèse, nous avons établi un guide d'annotation pour l'anonymisation de documents cliniques. L'intégralité de ce guide d'annotation est donné en annexe A. Une précision importante relative à ce guide concerne le fait que ce guide a été établi tardivement, après que de premières expériences d'anonymisation automatique ont été réalisées sur le corpus de cardiologie. [1] En conséquence, ce guide a uniquement été utilisé pour produire un corpus de référence. Autrement dit, les principes d'anonymisation présentés dans le guide ont été définis dès le début du projet *(après étude des caractéristiques du corpus et des besoins du projet)*, mais ils n'ont fait l'objet d'une formalisation dans un guide d'annotation que récemment.

Nous avons produit ce guide d'annotation en nous fondant sur les dix-huit catégories utilisées par le HIPAA américain et en adaptant ces catégories aux caractéristiques du corpus utilisé *(en l'occurrence le corpus Akenaton en cardiologie)* ainsi qu'aux spécificités culturelles françaises. En effet, chaque guide d'annotation se doit de prendre en compte les spécificités du corpus étudié et les types de documents traités. [Mayer et al., 2009] ont ainsi défini un guide d'annotation tenant compte des caractéristiques du corpus Veterans Affairs, de manière à anonymiser les neuf types

1. Le guide d'annotation pour l'anonymisation a été rédigé en février 2012 alors que le système d'anonymisation par méthodes symboliques « Medina » a été produit préalablement, entre novembre 2008 et février 2012.

de documents qui composent ce corpus. Le guide d'annotation que nous avons défini a servi deux objectifs : (i) servir de référence pour l'annotation manuelle du corpus, et (ii) fixer les caractéristiques des systèmes appliqués sur le corpus.

Principes généraux

L'annotation est réalisée en utilisant des balises XML qui encadrent les informations à anonymiser. Chaque balise renvoie au type sémantique de l'information annotée. Chaque annotation doit être réalisée avec un empan maximal *(si deux tokens successifs appartiennent à la même catégorie — deux prénoms consécutifs, ou un numéro puis un nom de rue —, alors ils seront regroupés dans la même portion)*, à l'exception des cas suivants. Nous établissons le principe que le texte d'origine doit être préservé. Si deux tokens d'une même catégorie sont situés en fin de ligne pour le premier et en début de ligne pour le deuxième, alors les deux tokens seront annotés séparément. Les signes de ponctuation qui suivent une entité mais qui n'en font pas partie *(une entité suivie d'une virgule)* ne sont pas intégrés dans la portion. À l'inverse, une ponctuation constituante d'une entité *(les points d'une abréviation, le séparateur dans une date, les traits d'union dans les noms composés, etc.)* seront intégrés dans la portion annotée. Ces conventions d'annotation vont de paire avec une segmentation du texte en tokens que les annotateurs doivent respecter.

Catégories

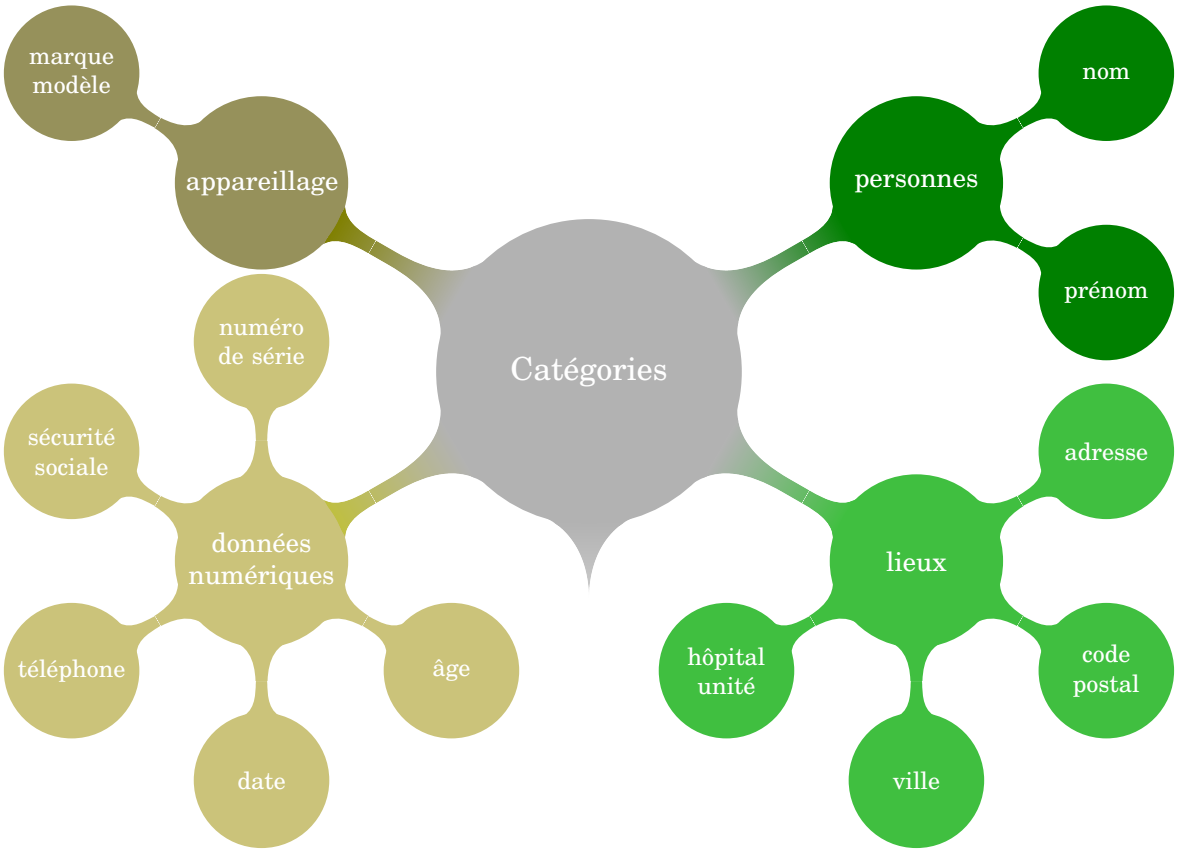


FIGURE 4.1 – Catégories couvertes par le guide d'annotation

Nous avons défini quatre principales catégories d'information, la plupart d'entre elles s'accompagnant de sous-catégories (figure 4.1).

Personnes. La première catégorie renvoie aux noms de personne et comprend deux sous-catégories : (i) les prénoms, qu'ils concernent les patients, les membres de la famille, ou le personnel hospitalier ; les déclencheurs *(M., Mme, Melle, etc.)* ne sont pas intégrés dans la portion annotée ; et (ii) les noms de famille, avec les mêmes caractéristiques et exceptions que pour les prénoms.

> Professeur [prénom H.] [nom MARTIN]
> M. [prénom Oscar] [nom Le Blanc]
> M. [nom Victor] [prénom Jérémie]
> Madame [nom Roux] [prénom Audrey]

Lieux. La seconde catégorie regroupe les informations de lieu et compte quatre sous-catégories : (i) les noms d'hôpitaux, de services ou d'unités dans un hôpital, les maisons de repos, etc. ; les déclencheurs *(Hôpital, Clinique, etc.)* sont intégrés dans la portion annotée ; (ii) les adresses physiques en tenant compte de tous les éléments constitutifs *(numéro de rue, boîte postale, nom de rue, etc.)*, que cette adresse soit celle du patient ou de l'hôpital ; (iii) tous les codes-postaux ; et (iv) tous les noms de ville sauf ceux figurant dans le nom d'un hôpital *(CHU de Rennes)* ; dans ce cas, le nom sera intégré dans la portion annotant l'hôpital.

> [hôpital CENTRE HOSPITALIER UNIVERSITAIRE DE RENNES]
> [hôpital Salle J.C. Dupont]
> Interne, [hôpital unité Ch. Durand]
> [hôpital Hôpital de Nantes]
> [adresse 9, rue de Saint-Malo] [codepostal 35000] [ville Rennes]

Données numériques. La troisième catégorie concerne cinq types de données numériques (hors numéros constitutifs d'une adresse postale) : (i) toutes les dates absolues ; les dates relatives *(le mois prochain)* et les durées *(pendant 3 mois)* ne sont pas annotées ; (ii) les âges dont la valeur est supérieure ou égale à 90 ans ; (iii) les numéros de série d'un appareillage médical ; (iv) le numéro de sécurité sociale du patient ; et (v) les numéros de téléphone et de télécopie. Les intervalles de dates sont regroupés en une seule portion *(en réduisant au maximum la présence de mots outils)* tandis que les conjonctions et disjonctions de dates seront séparées en deux portions.

> âgé de [âge 92 ans]
> né le [date 1er mars 2004]
> ([date 25/06])
> [date 1.1.1925]
> du [date 15 au 18 mars]
> en [date juin] et en [date septembre 2005]
> [numéro_ss 2 99 01 99 999 000]
> Rendez-vous au [téléphone 01.30.00.00.00 ou 01]

Appareillages. La dernière catégorie se rapporte aux appareillages médicaux et comprend essentiellement le nom de marque et le modèle.

> info St Jude Médical Microny II SR + modèle info 25-25 T (impédance de sonde à 750 ohms)
> STIMULATEUR : info Ela Medical BRIO DR 212
> Marque : info CPI Type : info 1000 No de série : numéro 1234567890

4.3 Les corpus

L'intégralité des expériences d'anonymisation présentées dans ce manuscrit porte sur un corpus de documents en cardiologie, constitué à l'occasion du projet Akenaton.

4.3.1 Présentation du projet Akenaton

Le projet Akenaton s'inscrit dans le domaine de la cardiologie, et plus spécifiquement, celui de la télécardiologie. La télécardiologie consiste en partie à transmettre quotidiennement au médecin traitant les alertes médicales générées par le pacemaker implanté chez un patient. Devant l'afflux d'informations que peut produire un tel type d'appareillage, il est nécessaire de rationaliser les alertes reçues. Ainsi, une alerte de même intensité chez deux patients différents n'aura pas le même impact ni les mêmes conséquences selon qu'elle concerne l'un ou l'autre des deux patients. Une rationnalisation automatique des alertes doit pouvoir être effectuée, sur la base des éléments constituant le dossier médical du patient, de manière à trier les alertes selon leur niveau de gravité afin que le médecin traitant puisse prendre les bonnes décisions sans se laisser déborder. L'objectif global poursuivi par ce projet consiste à traiter les alertes transmises quotidiennement par les défibrillateurs cardiaques dans un cas de figure bien précis : celui du risque de thrombo-embolie pulmonaire chez les patients victimes de fibrillation atriale.

L'un des éléments clés en matière de prévention des risques d'attaque thromboembolique chez des patients avec fibrillation atriale est le score CHA2DS2-VASc,[2] proposé par la Société européenne de cardiologie. Ce score est composé de huit critères qui comptent chacun pour un à deux points, avec un maximum de neuf points :

1. insuffisance cardiaque ou dysfonction ventriculaire gauche : 1 pt,
2. hypertension : 1 pt,
3. âge$\geq$75 ans : 2 pts,
4. diabètes : 1 pt,
5. accident vasculaire cérébral, attaque ischémique transitoire, ou embolie périphérique : 2 pts,
6. pathologie vasculaire *(infarctus du myocarde, athéromatose aortique en plaque, artérite périphérique)* : 1 pt,
7. 65$\leq$âge$<$75 : 1 pt,
8. et sexe féminin : 1 pt.

2. La signification de l'acronyme repose sur les critères énoncés en anglais : Congestive heart failure, Hypertension, Age 2 pts, Diabetes mellitus, Stroke 2 pts, Vascular disease, Age, Sex category.

Le type de traitement à administrer est fonction du score calculé [Olesen et al., 2012, Lip, 2013]. Un score nul n'appelle aucun traitement. Pour un score de un point, un traitement à base d'anticoagulant oral *(héparine, warfarine, anti-vitamine K, etc.)* est recommandé. Enfin, le traitement par anticoagulant oral est obligatoire pour un score supérieur ou égal à deux points.

Afin de mettre au point le système d'extraction d'information qui permettra d'instancier les critères de ce score, une étape préalable d'anonymisation automatique des comptes rendus s'est donc révélée obligatoire pour pouvoir sortir les dossiers patients du CHU partenaire.

4.3.2 Processus d'anonymisation poursuivi

Problématique

L'une des problématiques en matière de développement d'outils d'anonymisation de corpus médicaux concerne l'accès aux documents médicaux tout en assurant le respect de la vie privée du patient exposée dans ces documents. La question qui doit être résolue concerne le moyen de mettre à disposition des corpus médicaux composés de données réelles sans que les données personnelles ne soient présentes (voir section 1.5.1). La disponibilité de ce type de données permet de mettre au point des méthodes d'anonymisation sur des textes réels sans pour autant travailler sur des données fortement identifiantes.

Anonymisation à la source

La solution que nous avons mise en place dans le projet Akenaton consiste à effectuer une première passe d'anonymisation reposant sur des méthodes simples, qui peuvent être mises en œuvre à la source lors de la collecte des textes, mais permettant de supprimer la quasi-totalité des informations identifiantes fondamentales : le nom, le prénom et la date de naissance du patient. [3]

Ces informations personnelles figurent dans la partie structurée du dossier patient. Elles ont été extraites par Olivier Dameron depuis le fichier de métadonnées associé à chaque document textuel. Ces métadonnées contiennent en tout :

- données patient : nom, nom marital, prénom, date de naissance et sexe ;
- données de provenance du document : identifiant de l'unité fonctionnelle (UF), date de création et nom de l'auteur.

L'algorithme d'anonymisation à la source a été conçu pour réaliser les opérations suivantes, au sein de l'entrepôt de données hospitalier :

- remplacer le texte mentionnant le nom du patient, le nom martial et le prénom (quelle que soit la casse) respectivement par une balise <Nom patient>, <Nom marital patient> et <Prenom patient> ;
- remplacer la date de naissance par une balise <Date naissance patient>, d'après une identification basée sur trois types de motifs : jj/mm/aa, jj/mm/aaaa et 01 Janvier 1970.

Pour chaque document, le nombre de remplacements a été comptabilisé et les documents pour lesquels aucun nom, prénom ou nom marital n'a été trouvé ont été vérifiés manuellement.

3. L'ensemble de ce travail d'anonymisation de premier niveau a été réalisé par Olivier Dameron (INSERM U936 *Modélisation Conceptuelle des Connaissances Biomédicales*, Université de Rennes 1) au sein du CHU Pontchaillou à Rennes d'où provient le corpus.

Chiffrement des méta-données

Enfin, dans les fichiers de métadonnées, les noms, prénom et date de naissance du patient et le nom du médecin ont été chiffrés selon l'algorithme SHA-256, et les autres informations supprimées. [4] Ces fichiers, fournis avec le corpus, conservent donc une information permettant de mettre en évidence les documents parlant d'un même patient, mais sans possibilité de remonter à l'identité du patient. Cette méthode est similaire à celle employée par [Quantin et al., 2005] pour créer des identifiants familiaux chiffrés dans le cadre du partage des informations de santé entre États européens.

L'ensemble des documents (texte et métadonnées résiduelles codées) constitue le corpus à l'issue de cette première passe d'anonymisation à la source. C'est sur la base de ce corpus anonymisé de manière minimale que nous avons développé nos méthodes d'anonymisation présentées dans les chapitres suivants (chapitre 5 pour les méthodes symboliques et chapitre 6 pour les méthodes par apprentissage statistique).

4.3.3 Constitution des corpus annotés

Le corpus du projet Akenaton se compose de 21 749 documents médicaux, provenant du service de cardiologie du CHU Pontchaillou à Rennes. [5] Ces documents se rapportent à 11 964 patients différents et relèvent essentiellement de lettres de sortie établies par le chirurgien à destination du médecin traitant *(ce courrier reprend les principaux points de l'intervention chirurgicale, le traitement de sortie, et les observations éventuelles, voir section 1.2.1)*. L'ensemble du corpus a fait l'objet d'une anonymisation à la source telle que précédemment décrite.

De manière à évaluer le résultat des différentes expériences d'une part, et à fournir aux algorithmes d'apprentissage statistique une base de documents annotés d'autre part, nous avons constitué un corpus annoté rassemblant un sous-ensemble réduit du corpus global. Nous avons d'abord extrait aléatoirement 312 documents sur les 21 749 du corpus global *(utilisés lors des expériences à base de méthodes symboliques)*, puis, quelques mois plus tard, 250 nouveaux fichiers pour compléter la première extraction *(utilisés pour fournir plus de contextes d'apprentissage à l'outil d'apprentissage statistique)*, soit un total de 562 fichiers répartis en trois sous-ensembles. Le corpus se composant de documents textuels non annotés, aucun échantillonnage de corpus [6] fondé sur les catégories d'entités n'a été réalisé pour constituer cet ensemble de fichiers. Nous n'avons pas non plus effectué d'échantillonnage sur des catégories plus générales, n'ayant pas connaissance des types de documents d'origine *(compte rendu opératoire, lettre de sortie, etc.)* ou des auteurs *(parmi plusieurs chirurgiens)*.

4. Le script de chiffrement a été mis au point par Pierre Zweigenbaum (LIMSI–CNRS).

5. http://www.chu-rennes.fr/

6. Échantilloner un corpus revient à sélectionner dans les documents de base un sous-ensemble représentatif des caractéristiques à étudier. En ce qui concerne des tâches de repérage d'entités nommées ou d'anonymisation automatique, l'échantillonnage revient à sélectionner des documents dont le contenu permettra de s'assurer que chaque catégorie d'entité nommée est composée d'un nombre minimum d'occurrences. Un échantillonnage peut également être fondé sur des caractéristiques générales ne relevant pas des catégories d'entités à traiter : type de document, taille des documents, etc.

Réintroduction de données nominatives

Puisque le corpus mis à notre disposition a été anonymisé à la source sur les données patients, et parce que nous souhaitons évaluer tous les types d'entités, nous avons automatiquement réintroduit des noms et des prénoms dans le sous-corpus aléatoirement constitué à la place des balises de premier niveau <Prenom patient>, <Nom patient> et <Nom marital patient>. Cette réintroduction a été réalisée sur la base de deux listes de noms et de prénoms : (i) une première liste de noms et prénoms francophones que nous utilisons également dans nos expériences d'anonymisation, et (ii) une seconde liste internationale qui ne fait pas partie des ressources utilisées dans nos expériences. Nous avons utilisé le contenu de ces listes selon un ratio de 75 % de noms et prénoms francophones et 25 % de noms et prénoms internationaux. De cette manière, tous les noms et prénoms réintroduits ne sont pas directement identifiables par la projection de la liste utilisée dans nos expériences.

Au final, nous disposons d'un corpus respectant la vie privée du patient *(les données personnelles ne sont plus celles d'origine)* mais qui se compose néanmoins d'informations nominatives et numériques qu'il importe de traiter dans le processus d'anonymisation.

Processus d'annotation

Étapes d'annotation. Le corpus de 562 fichiers rassemblés pour créer un corpus annoté a été réalisé en trois étapes (voir figure 4.2).

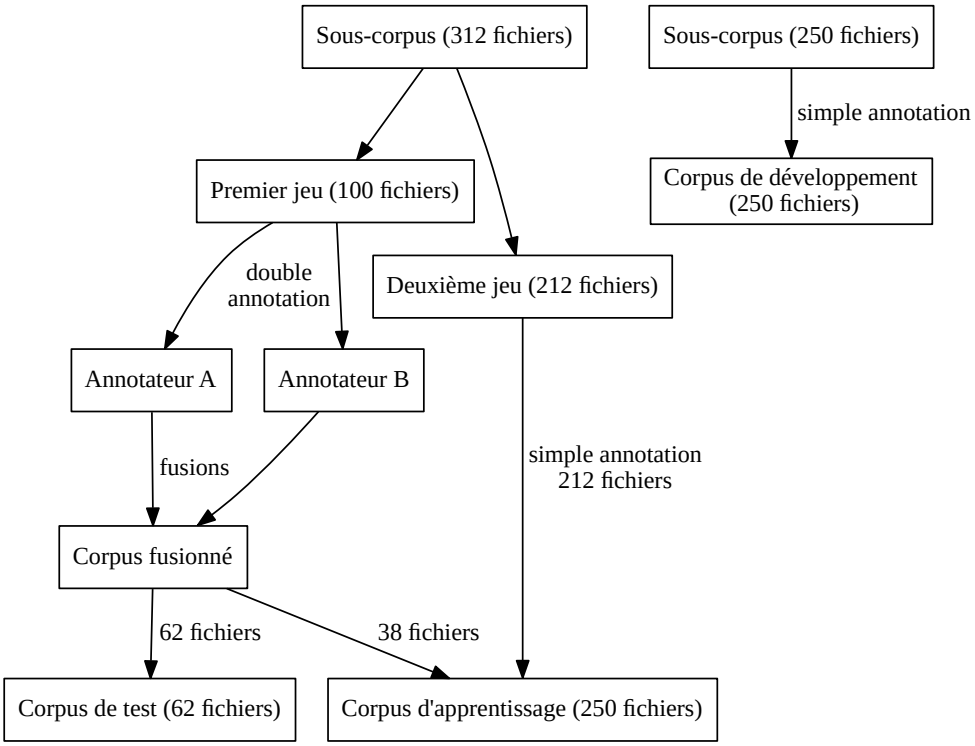


FIGURE 4.2 – Processus de constitution des corpus annotés

Dans une première étape, 100 fichiers ont été annotés en double [7] sur la base du guide d'annotation précédemment défini (voir section 4.2). Ces annotations ont ensuite fait l'objet d'une fusion en comparant les annotations produites par les deux annotateurs. Le résultat de cette fusion aboutit à des corrections, des suppressions et des ajouts d'annotations selon les cas. Sur la base des 100 fichiers résultants de la fusion, nous avons décidé de constituer le corpus de test intégralement avec des fichiers provenant de cet ensemble annoté en double, de manière à assurer une meilleure qualité d'annotation pour évaluer les différents systèmes : 62 fichiers ont donc été utilisés pour le corpus de test, les 38 fichiers restants ayant servi pour le corpus d'apprentissage.

Dans une seconde étape, 212 fichiers supplémentaires ont été annotés en simple annotation en cinq heures, puis corrigés par la même personne quelques jours plus tard en cinq heures également. Ces 212 fichiers ont intégralement été versés dans le corpus d'apprentissage, en complément des 38 fichiers annotés en double, constituant un corpus d'apprentissage global composé de 250 fichiers.

Dans une troisième étape, 250 fichiers nouveaux fichiers ont été annotés et corrigés par un seul annotateur, pour constituer le corpus de développement utilisé lors des expériences d'apprentissage statistique.

Nous renseignons dans le tableau 4.1 le nombre et le pourcentage d'entités annotées dans chaque corpus pour chaque catégorie d'entités. Le pourcentage d'entités apparaît également sur le graphique 4.3.

Catégorie	Corpus d'apprentissage	Corpus de développement	Corpus de test
Dates	819 (33,1 %)	829 (31,1 %)	238 (36,4 %)
Noms	845 (34,2 %)	937 (35,2 %)	205 (31,3 %)
Prénoms	454 (18,4 %)	505 (18,9 %)	109 (16,7 %)
Hôpitaux	173 (7,0 %)	161 (6,0 %)	43 (6,6 %)
Villes	84 (3,4 %)	94 (3,5 %)	22 (3,4 %)
Codes postaux	13 (0,5 %)	16 (0,6 %)	8 (1,2 %)
Adresses	11 (0,4 %)	12 (0,5 %)	8 (1,2 %)
Téléphones	38 (1,5 %)	76 (2,9 %)	8 (1,2 %)
Appareillage	26 (1,1 %)	22 (0,8 %)	10 (1,5 %)
Numéro de série	10 (0,4 %)	13 (0,5 %)	3 (0,5 %)
Nombre total d'entités	2 473	2 665	654
Nombre total de documents	250	250	62
Nombre moyen d'entités par document	9,89	10,66	10,55

TABLE 4.1 – Nombre et pourcentage d'entités dans chaque catégorie dans les corpus

Cohérence des annotations. La cohérence des annotations — le caractère régulier des annotations produites — a été assurée de trois manières distinctes et complémentaires.

7. L'annotation a été réalisée par Louise Deléger (*Cincinnati Children's Hospital Medical Center — Division of Biomedical Informatics*, Cincinnati, OH) et par moi-même. Autrement dit, au moins une personne extérieure à la production du guide d'annotation, Louise n'ayant pas participé à la conception du guide.

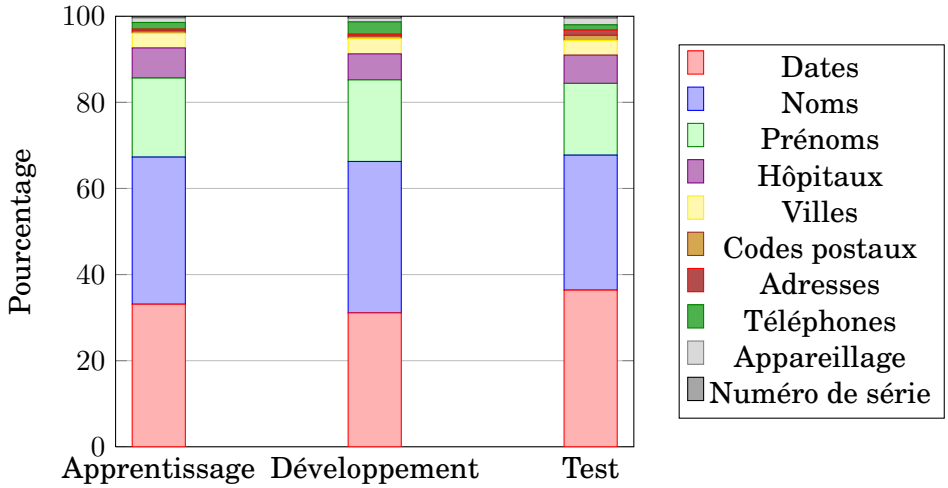


FIGURE 4.3 – Pourcentage d'entités de chaque catégorie dans les corpus

En premier lieu, en nous référant en permanence au guide d'annotation. Bien que connaissant le guide d'annotation pour l'avoir rédigé, certains points de détail peuvent nous échapper *(notamment sur l'intégration des ponctuations*[8] *et des déclencheurs*[9] *dans les portions annotées)*. La consultation systématique du guide d'annotation permet de garantir une cohérence dans les annotations réalisées.

Un second moyen de nous prémunir contre l'absence de cohérence dans les annotations consiste à utiliser des outils pour détecter automatiquement ces incohérences. L'approche que nous avons mise en place consiste à appliquer nos outils d'anonymisation sur le corpus pour lequel nous venons de produire des annotations manuelles, puis à étudier les sorties produites par les systèmes notamment en cas de désaccord. Cette étude permet de mettre en évidence : (i) des entités oubliées lors de l'annotation *(oubli par mégarde ou par incompréhension d'une règle du guide)* et (ii) des erreurs de frontières dans les portions annotées *(généralement l'intégration d'un mot outil ou d'une ponctuation dans la portion annotée)*. Dans cette optique, les outils d'anonymisation sont utilisés comme des aides à l'annotation et peuvent être vus comme des outils permettant une pré-annotation automatique. L'objectif de cette approche ne consiste pas à produire une référence qui correspond aux sorties des systèmes que l'on souhaite évaluer par la suite *(ce qui conduirait à un biais évident)*, mais bien à s'inspirer des annotations systématiques réalisées par les outils automatiques pour contrer l'absence de systématicité inhérente aux annotations humaines.

Enfin, le dernier moyen utilisé revient à vérifier une seconde fois les annotations produites au moyen d'un nouveau passage manuel sur chaque fichier, éventuellement complété par l'utilisation de petits scripts.[10] Même si cette démarche de vérification est coûteuse, elle se révèle néanmoins plus rapide que la première phase d'annotation lorsque le texte n'est encore porteur d'aucune annotation.

8. Le guide d'annotation précise que les ponctuations doivent être sorties des portions annotées, sauf si elles sont constitutives d'une abréviation dans la portion. Le point dans les initiales et abréviations sera conservé tandis que les ponctuations encadrant les entités figureront hors de la portion.

9. Les déclencheurs de noms de personne ne sont pas intégrés dans la portion annotée alors que les déclencheurs d'hôpitaux le sont.

10. Un script projetant les douze mois de l'année sur les fichiers annotés permet, par exemple, de vérifier que les dates intégrant les mois en toutes lettres ont bien été annotées.

Accords inter-annotateurs. Nous avons calculé le taux d'accord inter-annotateurs entre les deux ensembles de 100 fichiers annotés en double au moyen de l'outil *ne-scoring-gen* (voir section 4.4.2) développé par Olivier Galibert. Nous donnons dans le tableau 4.2 les scores calculés par cet outil au moyen des différents coefficients, même si seul le coefficient κ fait sens dans le cadre de nos travaux pour les raisons présentées dans le chapitre 3.4.

Avant fusion des annotations, nous observons que l'accord inter-annotateur calculé entre les annotateurs A et B ($\kappa = 0,8073$) se situe au niveau du seuil de 0,8 retenu par [Artstein et Poesio, 2008] pour considérer qu'il y a un accord entre annotations. Nous avons également calculé les taux d'accords inter-annotateur entre chaque annotateur et le résultat de la fusion. Nous observons ainsi qu'il y a un meilleur accord entre les annotations réalisées par l'annotateur B et la fusion ($\kappa = 0,9307$) qu'entre les annotations de l'annotateur A et la fusion ($\kappa = 0,8698$).

	A vs. B	A vs. fusion	B vs. fusion
Coefficient S	0,8493	0,8992	0,9464
Coefficient π	0,8072	0,8697	0,9307
Coefficient κ	0,8073	0,8698	0,9307
Vrais Positifs	834	879	923
Faux Positifs	110	91	47
Faux Négatifs	97	52	21
F-mesure	0,8896	0,9248	0,9645
Intervalles de confiance sur la F-mesure *(simulation de Monte Carlo pour* $n = 10^8$*)*	[0,8747;0,9045]	[0,9125;0,9371]	[0,9561;0,9729]

TABLE 4.2 – Taux d'accords inter-annotateurs sur les 100 fichiers annotés en double et intervalles de confiance sur la F-mesure

Discussion. Le résultat des accords inter-annotateurs ainsi que les différences de significativités statistiques sont étayés par le fait que pendant la phase d'adjudication des résultats produits par les deux annotateurs, nous avons constaté que les annotations de l'annotateur B étaient plus souvent correctes que celles de l'annotateur A. Cette observation empirique se traduit donc, pour le couple « A vs. fusion », par un taux d'accord inter-annotateur plus faible d'une part, et par une différence statistique réelle d'autre part. Les conclusions sont inversées pour l'accord inter-annotateur et la significativités calculés sur le couple « B vs. fusion ».

4.3.4 Utilisation des corpus annotés

Ce corpus de 562 fichiers annotés est utilisé de manière complémentaire dans les différentes expériences d'anonymisation que nous avons menées. Nous résumons ci-dessous l'usage des fichiers composant ce corpus tel qu'il a été envisagé lors des constitutions :

– le corpus d'apprentissage a été utilisé (i) pour définir les règles *(patrons syntaxiques)* des outils reposant sur les méthodes symboliques, et (ii) comme base annotée pour construire le modèle dans les méthodes par apprentissage statistique ;

– le corpus de test a été utilisé pour évaluer les sorties des outils relevant des différentes approches. Il n'a fait l'objet d'aucune consultation lors de la réalisation des différentes expériences. Parce qu'il est utilisé dans les évaluations, nous avons souhaité nous assurer de sa qualité. L'intégralité des 62 fichiers qui composent ce sous-corpus provient du processus de fusion de la double annotation ;
– le corpus de développement a été constitué ultérieurement aux deux autres, [11] lors de la réalisation des expériences par apprentissage statistique. Les premières expériences ont en effet démontré que le corpus d'apprentissage n'était pas suffisant, en termes de représentation des contextes d'apparition des différentes informations, et qu'un corpus complémentaire était donc nécessaire pour assurer au modèle une meilleure robustesse. Ce corpus est donc uniquement utilisé dans les expériences par apprentissage, comme moyen d'optimisation des prédictions du modèle.

Nous rappelons schématiquement dans le tableau 4.3 les modalités de constitution des différents sous-corpus, en termes de nombre de fichiers issus de chaque ensemble d'extraction *(trois jeux de 100, 212 et 250 fichiers)* et de répartition de ces fichiers parmi les trois sous-corpus envisagés *(apprentissage, développement et test)*.

Jeu d'extraction	Premier		Deuxième	Troisième
Nombre de fichiers	100		212	250
Type d'annotation	Double		Simple	Simple
Nombre de fichiers	62	38	212	250
Sous-corpus d'affectation	Test	Apprentissage		Développement
Nombre de fichiers	62	250		250

TABLE 4.3 – Modalités de constitution des sous-corpus utilisés dans les différentes expériences : apprentissage, développement, test

4.4 Les outils utilisés et développés

Dans cette section, nous présentons très succintement les outils que nous avons utilisés. Ces outils sont répartis en trois catégories : (i) les outils d'annotation de corpus, pour permettre la constitution des corpus de référence, (ii) les outils d'évaluation des annotations produites par les humains ou par les systèmes, et (iii) les outils d'interprétation des résultats produits. Pour chacune de ces catégories, nous avons utilisé des outils existants ou des fonctions déjà implémentées.

4.4.1 Annotation de corpus

L'annotation de corpus consiste à appliquer des annotations *(généralement les informations que le système devra identifier automatiquement)* sur un corpus de documents.

11. Les corpus d'apprentissage et de test ont été produits aux mois de février et mars 2012. Le corpus de développement a été produit en décembre 2012.

Types d'annotations. Les annotations peuvent être embarquées *(elles sont directement apposées sur le corpus)* ou débarquées *(elles figurent dans des fichiers distincts de ceux contenant les textes étudiés)*. L'annotation embarquée présente l'avantage de stocker les informations dans le même fichier que le contenu et autorise ainsi un accès direct à l'information, facilitant la lecture pour un humain. Elle ne permet cependant pas le chevauchement de balises *(pour le cas où deux portions annotées se superposent partiellement)* et imposent de définir un format pivot si plusieurs outils sont amenés à travailler sur les données annotées. L'annotation débarquée est stockée dans un fichier annexe au document étudié. Elle présente l'avantage de préserver le texte d'origine au caractère près et permet la gestion d'informations complexes de plusieurs types. Cependant, elle induit un manque de lisibilité pour un humain et suppose d'utiliser des outils spécifiques pour visualiser les textes annotés et modifier les annotations *(les références des annotations étant généralement indiquées au moyen d'une position dans le texte en terme de numéro de caractère ou de numéro de token)*.

Pour ces questions de lisibilité, nous avons choisi de recourir aux annotations embarquées, en retenant le formalisme XML pour encadrer les informations à anonymiser de balises indiquant le type d'information traité.

Outil d'annotation. Pour ce travail d'annotation de corpus, nous avons utilisé le module d'annotation développé par Olivier Galibert dans le cadre du projet Quaero pour l'éditeur de textes Xemacs. Ce module offre de nombreux raccourcis clavier configurables [12] pour faciliter le travail d'annotation [Grouin et al., 2011]. Ayant déjà utilisé ce module pour annoter des corpus, celui-ci s'est révélé d'un usage aisé pour un résultat adapté à notre tâche.

4.4.2 Évaluation des résultats

Afin d'évaluer la qualité des annotations produites par les humains d'une part *(lors de l'annotation de corpus)*, et la qualité des résultats générés par les différents systèmes d'autre part, nous avons utilisé un outil d'évaluation générique, également développé par Olivier Galibert pour le projet Quaero [Galibert et al., 2011]. Intitulé *ne-scoring-gen*, cet outil permet de calculer deux types de mesures.

Le premier type renvoie aux mesures habituellement utilisées dans les tâches de classification automatique *(rappel, précision, F-mesure)*, pour chacune des classes utilisées et de manière globale, complétées par la mesure composite du *Slot Error Rate* qui évalue à la fois le typage et les frontières d'une annotation (voir section 3.2).

Le deuxième type de mesure concerne les accords inter-annotateurs entre deux annotateurs (voir section 3.4). L'outil implémente les algorithmes de calcul de trois coefficients (S, π, κ) avec la possibilité facultative de spécifier le nombre de marquables [13] sur lesquels se fonder pour réaliser ces accords. Les calculs que nous pré-

12. L'annotation d'une entité est réalisée en sélectionnant le mot à annoter — ou en positionnant le curseur sur le mot — suivi de la combinaison de touches F2 + l'initiale du type d'entité à représenter *(par exemple, F2+p pour un prénom, F2+n pour un nom, etc.)*. La touche F5 permet de supprimer l'annotation sur laquelle se situe le curseur. Des fonctionnalités supplémentaires, utiles pour Quaero, sont également proposées par cet outil *(annotation de composants transverses ou spécifiques d'entités, correction de la forme sous laquelle apparaît l'entité, etc.)*, mais nous n'en avons pas eu l'usage.

13. Nous appelons « marquables » les unités d'un texte qui peuvent être annotées. Dans le cadre d'une annotation en entités nommées, il est difficile, voire impossible, de déterminer le type des unités qui peuvent être annotées *(les mots, les tokens, les syntagmes ?)*. La comparaison du nombre de marquables

sentons dans les chapitres suivants ont été réalisés sans spécifier le nombre de marquables.

4.4.3 Interprétation des résultats

Une fois obtenus les résultats des annotateurs humains ou des systèmes, il importe de comparer ces résultats entre eux de manière à interpréter la validité de ces résultats (voir section 3.5). Nous avons appliqué les deux approches généralement utilisées au moyen de fonctions sous le logiciel R.

Enfin, nous avons procédé au calcul des intervalles de confiance dans lesquels se situe chaque F-mesure au moyen d'une fonction R développée au CCHMC [14] reproduisant la simulation de Monte Carlo (voir section 3.5.2). Cette fonction repose sur les valeurs de vrais positifs, faux positifs et faux négatifs d'une expérience donnée, avec n tirages au hasard pour un risque α donné. Nous avons retenu la valeur $n = 10^8$ et un risque $\alpha = 0,05$ pour les différentes simulations réalisées.

4.5 Synthèse

Guide d'annotation. Un guide d'annotation est un manuel destiné à des annotateurs humains pour annoter un corpus. Il présente le corpus et l'objectif poursuivi, les règles d'annotation et un grand nombre d'exemples permettant à l'humain de se familiariser avec la tâche d'annotation. Le corpus ainsi annoté par des humains fait ensuite l'objet d'adjudications et de fusion des annotations pour disposer d'un corpus final de qualité. Ce corpus est alors utilisé, soit pour évaluer les performances d'un système, soit comme base d'apprentissage pour les outils reposant sur des méthodes par apprentissage statistique. Le guide d'annotation ne doit pas être confondu avec le cahier des spécifications fonctionnelles attendues. Tous les cas de figure ne sont pas référencés dans le guide d'annotation et les règles qui y sont précisées ne sont pas exhaustives.

Sur la base de ces caractéristiques, nous avons établi un guide d'annotation pour l'anonymisation. Quatre catégories principales d'information ont été définies avec des sous-catégories : (i) personnes *(noms et prénoms)*, (ii) lieux *(hôpitaux, adresses physiques, codes postaux, villes)*, (iii) données numériques *(dates absolues, relatives et durées, âges≥90 ans, numéros de série, de sécurité sociale, de téléphone)*, et (iv) marque et modèles des appareillages médicaux. Nous avons illustré chaque règle d'exemples fictifs ou réels.

Corpus. Nous avons utilisé un corpus de 21 749 documents médicaux dans le domaine de la cardiologie provenant du projet Akenaton pour réaliser nos expériences d'anonymisation automatique. De ce corpus global, nous avons aléatoirement extrait 100 fichiers qui ont fait l'objet d'une annotation manuelle par deux annotateurs humains, permettant l'obtention d'un corpus annoté de qualité. Le taux d'accord inter-annotateurs se monte à 0,8073 *(coefficient κ)* entre les annotations produites. Ce corpus annoté a ensuite fait l'objet d'une étape d'adjudication des résultats. Des accords inter-annotateurs ont été calculés entre l'annotation humaine d'origine et le résultat de cette fusion, avec des coefficients κ qui varient entre 0,8698 pour le premier annotateur et 0,9307 pour le deuxième.

utilisés dans le calcul des accords inter-annotateurs a été décrite dans [Grouin et al., 2011].

14. Cincinnati Children's Hospital Medical Center, Cincinnati, OH.

Une deuxième étape d'extraction aléatoire de 212 fichiers a été réalisée, donnant lieu à une annotation manuelle de ces fichiers en simple annotation.

Le corpus manuellement annoté a été scindé en deux sous-corpus, avec un premier corpus de 62 fichiers *(issus du jeu de données annotées en double)* pour l'évaluation, et un deuxième sous-corpus de 250 fichiers *(composé des 38 fichiers restants des annotations en double et des 212 fichiers annotés en simple)* pour l'entraînement des systèmes à base d'apprentissage statistique.

Enfin, un troisième sous-corpus, utilisé comme corpus de développement pour les expériences à base d'apprentissage statistique, a été produit, sur la base d'une extraction aléatoire de 250 fichiers avec une annotation en simple.

La cohérence des annotations a été vérifiée de trois manières différentes : en consultant fréquemment le guide d'annotation, en utilisant les outils d'anonymisation existants comme moyen de détection d'erreurs de frontières ou d'oublis d'annotation, et en vérifiant chaque fichier annoté avec l'aide de scripts dédiés.

Outils. Nous avons réutilisé un certain nombre d'outils existants et fonctions mathématiques déjà implémentées. Le travail d'annotation de corpus a été réalisé au moyen d'un module Xemacs développé pour le projet Quaero, permettant une annotation embarquée au moyen de balises typantes XML. Le calcul des accords inter-annotateurs et des mesures de performance des systèmes *(rappel, précision, F-mesure, Slot Error Rate)* présentés au chapitre 3 a été obtenu via l'outil *ne-scoring-gen* également développé pour le projet Quaero. Enfin, le calcul des intervalles de confiance a été réalisé par le biais de fonctions sous le logiciel R.

Chapitre 5

Méthodes symboliques

Pour les citoyens d'Ankh-Morpork, l'orthographe était pour ainsi dire en sus. Ils y croyaient comme ils croyaient à la ponctuation ; peu importe où on la plaçait du moment qu'elle était là.

La Vérité
TERRY PRATCHETT

Sommaire

5.1 Introduction

Dans ce chapitre, nous présentons les différentes expériences à base de méthodes symboliques que nous avons menées. Nous avons ainsi réalisé deux expériences principales, la première reprenant un outil existant, la seconde se fondant sur la création d'un nouvel outil.

La première expérience consiste à reprendre un outil existant pour l'anglais, librement réutilisable et modifiable, de manière à l'adapter au français. Nous détaillons les différentes étapes d'adaptation effectuées et les difficultés que nous avons rencontrées, en particulier celles qui ne nous ont pas permis de pousser l'adaptation au français jusqu'à son terme.

Partant de ce constat, nous avons réalisé une seconde expérience qui consiste à produire un nouvel outil. Nous avons envisagé le fonctionnement de cet outil en reprenant les mêmes principes que ceux utilisés dans les outils existants. Nous présentons les principales caractéristiques de cet outil et les évaluations que nous avons menées.

Préalablement à la présentation de ces deux expériences, nous rappelons brièvement les premières approches de l'anonymisation automatique de documents cliniques que nous avons abordées lors d'un stage de fin d'études.

5.2 L'outil « Stomato » : premières approches de l'anonymisation

Dans cette section, nous présentons le travail de production d'un corpus médical anonymisé et l'outil d'anonymisation « Stomato » que nous avons produit à l'occasion d'un stage réalisé au sein de la mission de recherche en Sciences et Technologies de l'Information Médicale (STIM) au CHU de la Pitié-Salpétrière, sous la direction de Pierre Zweigenbaum. Ce stage de fin d'études a été effectué pour valider le DESS *Ingénierie Multilingue* de l'INaLCO. [1] L'ensemble du travail présenté dans cette section a été réalisé en 2002. Il constitue les premières approches que nous avons eues concernant la problématique de l'anonymisation de données médicales [Grouin, 2002].

L'objectif final de ce travail consistait à produire un corpus médical anonymisé pour l'intégrer dans un corpus plus vaste du français contemporain *(projet CLEF*[2]*).*

5.2.1 Corpus utilisé

Le corpus utilisé pour peupler le projet CLEF au titre du domaine médical provient de deux sources distinctes et rassemble, à ce titre, des documents de types et de contenus distincts.

- Établissement Français des Greffes (EFG), 134 documents répartis parmi quatre principaux types de documents *(arrêtés, articles du code de la santé publique, circulaires, et décrets)*, complétés par les cinq meilleure copies du concours national de philosophie sur le sujet « *Ce que je donne dans le don d'organes, est-ce une partie de moi-même ?* » (session de juin 2001) ;
- CHU de la Pitié-Salpétrière, service de stomatologie/chirurgie maxillo-faciale, 9 745 documents rédigés par neuf praticiens parmi trois types de documents

1. Institut National des Langues et Civilisations Orientales, Paris.
2. http://estime.spim.jussieu.fr/CLEF/, projet piloté par Didier Bourigault (ERSS), Benoît Habert (ENS Fontenay/Saint-Cloud), Patrick Paroubek (LIMSI-CNRS) et Pierre Zweigenbaum (AP-HP, DIAM, SIM).

(certificats médicaux initiaux, comptes rendus opératoires, et lettres de correspondance avec le médecin traitant).

Nous avons extrait de ce corpus global un sous-corpus en stomatologie pour lequel nous souhaitions garantir la qualité des anonymisations réalisées. La constitution de ce sous-corpus a été réalisée d'après le protocole suivant : pour chacun des neuf praticiens, nous avons aléatoirement extrait quinze documents de chacun des trois types de documents disponibles, soit un total envisagé de 405 documents. Deux dossiers ne contenant aucun certificat médical initial, le sous-corpus final compte 375 documents, pour un total de 87 811 mots.

5.2.2 Processus d'anonymisation

La démarche suivie pour anonymiser le corpus de stomatologie repose sur la combinaison enchaînée de deux outils : l'outil d'anonymisation « Stomato » que nous avons développé suivi du « Scrubber » développé par Patrick Ruch [Ruch et al., 2000]. La combinaison a été envisagée pour augmenter le nombre d'informations traitées et diminuer en conséquence le nombre de sous-anonymisations. L'application du Scrubber sur les sorties de Stomato a permis d'augmenter de 205 le nombre d'informations anonymisées, soit une augmentation de 4 % du nombre total d'informations traitées.

Représentation de sortie. Les deux outils remplacent l'information à anonymiser de deux manières différentes :

- Pour le Scrubber, en remplaçant chaque caractère alpha-numérique d'origine par un "x" générique et en reproduisant les différences de casse typographique dans le résultat produit : *M. Jean Dupont* devient *M. Xxxx Xxxxxx* ;
- Des balises XML typantes configurables *(fichier "balises.txt")* pour Stomato de la forme : `<name type="person" />` *(personnes)* ou `<date />` *(dates).*

Propriétés du script Stomato. Le script que nous avons développé permet de traiter huit catégories d'informations *(noms, prénoms, âges, dates, noms d'hôpitaux, adresses postales, numéros de téléphone et codes des actes médicaux).* Il repose sur les propriétés suivantes :

- Cinq listes d'entités nommées dont le contenu est projeté sur chaque document *(13 472 noms de famille, 12 431 prénoms, 39 000 villes, 175 pays et 109 noms complets ou partiels d'hôpitaux avec déclencheurs)* ;
- Une série de neuf déclencheurs pour les noms de personnes *(Mr, Mme, Melle, Pr, Dr, Monsieur, Madame, Docteur, Professeur)* ;
- Dix-sept règles principales dont certaines reposent sur l'étude du voisinage des informations déjà anonymisées *(permet l'anonymisation du prénom présent sous forme d'initiale devant un nom anonymisé)* ;
- Un second passage d'anonymisation est effectué en étudiant le contexte gauche des informations traitées lors du premier passage.

Constitution d'un corpus de référence. Nous avons vérifié l'ensemble des fichiers traités par cette procédure, de manière à corriger les erreurs et produire un corpus de référence qui soit diffusable. Sur 375 fichiers, deux seulement ne contenaient aucune erreur. Ce travail de vérification humaine a nécessité 6h35 de travail. Aucune évaluation des résultats produits par la combinaison du Scrubber et de Stomato sur cet ensemble de documents n'a été réalisée.

5.2.3 Évaluation

Pour évaluer les deux systèmes, nous avons constitué un mini corpus composé de dix documents *(trois certificats médicaux initiaux, quatre comptes rendus opératoires et trois lettres de correspondance)*. Nous donnons dans le tableau 5.1 le résultat de l'évaluation comparée des deux systèmes sur ce mini corpus de test.

	Scrubber	Stomato
Vrais positifs	75	117
Faux positifs	13	18
Faux négatifs	66	24
Rappel	0,5319	0,8298
Précision	0,8523	0,8667
F-mesure	0,6550	0,8478
Intervalles de confiance sur la F-mesure *(simulation de Monte Carlo pour $n = 10^8$)*	[0,5838;0,7262]	[0,8025;0,8932]

TABLE 5.1 – Évaluation du Scrubber et de Stomato sur un corpus de dix documents

Les intervalles de confiance calculés sur la base de ces résultats témoignent d'une marge de progression possible pour les deux outils. Ils démontrent surtout que les résultats calculés sur un aussi faible échantillon *(le mini corpus compte 141 entités à anonymiser)* donnent lieu à une variation potentielle importante que seule une évaluation sur un plus grand nombre de données permettrait de confirmer ou d'infirmer. Faute de disposer du corpus de 375 fichiers d'une part, et du Scrubber d'autre part, nous ne sommes pas en mesure d'effectuer une évaluation à plus grande échelle.

5.2.4 Discussion

Le Scrubber a permis l'anonymisation de certains types d'informations seulement *(noms, prénoms, dates)*. Sur d'autres catégories *(adresses postales et codage des actes)*, le système n'a pu traiter efficacement ces informations faute de disposer de règles adaptées. Enfin, deux dernières catégories *(noms des hôpitaux et fonction du praticien)* n'ont pas été traitées car elles ne constituent pas un type d'information à anonymiser en Suisse.

Nous avons par ailleurs constaté que les règles utilisées dans les deux systèmes devaient faire l'objet d'amélioration, sur trois cas particuliers :

- Des anonymisations incomplètes *(erreur de frontière)* : *<address /> de Landy* ;
- Des anonymisations inutiles *(erreur d'insertion)* : *remis en xxxxx propres* ;
- Une gestion difficile des formats des deux outils conduisant à des chevauchements d'annotations : *l'enfant Xxxxx type="person" />DUPONT Laurène*.

Un dernier point concerne les faux positifs *(sur-anonymisations)* qui, pour ce qui concerne le domaine de la stomatologie, concernent fréquemment des noms de procédures chirurgicales composés de noms de personne ou de lieux *(Glasgow initial, intervention de Caldwell Luc, disjoncteur de Tessier, plaque de Champy, crochet de Ginestet, canule de Montandon, davier de Rowe et Keeley, levier de Kinley, etc.)*. Nous avions estimé que l'utilisation d'anti-dictionnaire *(liste noire)* serait bénéfique pour pallier ces sur-anonymisations.

5.3 L'outil « De-ID » : adaptation de l'anglais au français

Puisqu'il n'existe aucun outil d'anonymisation librement utilisable pour le français qui repose sur des méthodes symboliques, nous avons cherché à reprendre un outil existant pour l'anglais, de manière à l'adapter au français. Notre choix s'est porté sur l'outil « De-ID » réalisé à la division des technologies et sciences de la santé du MIT [Neamatullah, 2006, Neamatullah et al., 2008]. Les raisons de ce choix sont doubles. En premier lieu, les auteurs autorisent sa réutilisation et son adaptation, ce qui convient au but que nous poursuivons, d'autant plus que l'évaluation de l'outil sur l'anglais est positive (rappel de 0,967 et précision de 0,749). En second lieu, l'outil a été écrit dans un langage de script que nous maîtrisons, le langage Perl, [3] ce qui permet d'être immédiatement opérationnel dans son adaptation au français.

Compte-tenu de la manière dont l'outil « De-ID » a été envisagée *(système à base de listes et de règles)*, son adaptation au français repose sur trois étapes principales : (i) l'adaptation des listes utilisées, (ii) l'adaptation des déclencheurs, et (iii) l'adaptation des règles [Grouin et al., 2009a].

5.3.1 L'adaptation des listes

Le premier point se révèle assez simple à résoudre. Le site internet de l'Association des Bibliophiles Universels du CNAM fournit gracieusement plusieurs listes de mots et dictionnaires, [4] parmi lesquels des listes de 12 437 prénoms et de 39 076 villes françaises. Toutes les listes proposées ne sont pour autant pas d'égale qualité. Si le dictionnaire des noms communs se révèle riche et varié (plus de 251 000 formes fléchies), la liste des villes françaises est désaccentuée, tout comme la liste des prénoms. De plus, cette dernière intègre une majorité de prénoms anglo-saxons. Il s'agit cependant de ressources utiles qu'il est possible d'adapter (en tentant une réaccentuation automatique), ou pour lesquelles il est peut être envisagé de contraindre les outils de traitement automatique des langues à s'adapter (en étant insensibles aux caractères accentués par exemple, même si cela appauvrit le processus). Nous avons par ailleurs complété cet ensemble de listes par nos propres ressources, soit en reprenant des listes précédemment constituées *(liste de 33 380 noms de villes)*, soit en créant de nouvelles ressources *(liste de 2 526 noms d'hôpitaux, cliniques, centres hospitaliers français, constituée à partir d'un annuaire de santé présent sur Internet* [5] *)*.

5.3.2 L'adaptation des déclencheurs

Le second point est également aisé et rapide à effectuer. Il existe peu de déclencheurs utilisés dans un système d'anonymisation. Leur adaptation d'une langue à une autre, mais également d'une culture à une autre ne pose donc aucun problème. Trois catégories de déclencheurs ont ainsi été traduites ou complétées : (i) les particules de noms de famille *(De, Mc, Van, etc.)*, (ii) les titres *(Mister, Doctor, Professor, etc.)*, et (iii) les expressions précédant un lieu *(lives in, resident of, comes from, etc.)*. Le logiciel De-ID reposant sur des listes de déclencheurs distinctes selon que le déclencheur est attendu avant ou après le mot à anonymiser, la contrainte de l'ordre des mots qui diffère en anglais et en français a ainsi pu être aisément contournée. En

3. Practical Extraction and Report Language, langage de script à base d'expressions régulières.

4. http://abu.cnam.fr/DICO/, Association des Bibliophiles Universels, CNAM Conservatoire National des Arts et Métiers, Paris.

5. http://www.sanitaire-social.com/, Annuaire Sanitaire et Social.

raison de l'ordre des mots différents en français et en anglais, certains déclencheurs présents dans la liste des suffixes en anglais ont été déplacés dans la liste des préfixes en français : les suffixes anglais *Street, Road, Blvd* deviennent les préfixes *Rue, Route, Bld* en français.

5.3.3 L'adaptation des règles

Le dernier point, relatif aux règles, se révèle particulièrement complexe à résoudre. Dans cet outil, les règles ont été implémentées sous la forme d'expressions régulières rédigées en langage Perl. Nous avons constaté que modifier les expressions régulières constituait la tâche la plus difficile. Nous avons ainsi identifié deux problèmes majeurs qui ne nous ont pas permis d'achever le processus d'adaptation de l'outil au français. En effet, cette adaptation dépend d'une part de la manière dont le logiciel a été conçu *(du point de vue de la manière de coder)*, et d'autre part de l'objectif visé *(en l'occurrence, convertir le logiciel pour une utilisation en français, sur un corpus de comptes rendus médicaux)*. Si l'adaptation des expressions régulières pour les informations numériques *(numéros de Sécurité Sociale, téléphone, dates, etc.)* ne pose aucun problème particulier, il en est tout autrement des informations nominatives. La réalisation d'expressions régulières étant fortement contextuelle, il ne nous a guère été possible d'adapter davantage les expressions existantes sans les reprendre à la base, ce qui se serait révélé très chronophage.

5.3.4 Évaluation

Nous avons évalué les performances de la version francisée de De-ID, au niveau d'avancement que nous avons atteint. Nous avons constitué un corpus de test composé de vingt-trois documents cliniques issus du corpus de cardiologie. Parallèlement à cette adaptation, nous avons commencé le développement de notre propre outil, intitulé « Medina » (voir section 5.4). Nous donnons dans le tableau 5.2 le résultat de l'évaluation comparée des deux systèmes sur ce corpus de test.

	De-ID francisé	Medina
Vrais positifs	108	138
Faux positifs	372	12
Faux négatifs	58	28
Rappel	0,6506	0,8313
Précision	0,2250	0,9200
F-mesure	0,3344	0,8734
Intervalles de confiance sur la F-mesure *(simulation de Monte Carlo pour* $n = 10^8$*)*	[0,2875;0,3812]	[0,8347;0,9122]

TABLE 5.2 – Évaluation de De-ID francisé et de Medina sur un corpus de 23 documents

5.3.5 Discussion

La consultation des fichiers anonymisés par la version arrêtée de De-ID francisé permet de constater qu'il restait une charge de travail importante à accomplir, à tous les niveaux : poursuite du nettoyage des listes d'entités et réécriture des expressions

régulières du script. Une analyse des erreurs produites nous permet de dresser les constats suivants.

Problème de définition des règles. Le principal problème, et de loin le plus complexe, concerne la compréhension du fonctionnement des règles d'une part, et les problèmes de définition des règles d'autre part.

Compréhension du fonctionnement des règles. Comprendre pour quelle raison certaines règles se déclenchent *(alors qu'elles ne devraient pas être appliquées)* constitue un point essentiel pour réduire le nombre de sur-anonymisations (exemple 14 ; la balise <Nom patient> provient de l'étape d'anonymisation à la source, voir section 4.3.2).

(14) initiales A nom-famille son admission dans le service, Monsieur <Nom patient> est en insuffisance cardiaque sévère dyspnéique au moindre effort.

Dans l'exemple 15, la ville de *Versailles* et le mois de *mars* ont tous deux été typés comme étant des lieux par l'application de règles. Ces deux termes sont présents dans la liste des villes [6] *(ce qui déclenche l'étiquetage)*, mais l'outil a indiqué un caractère d'incertitude sur le fait qu'il s'agit réellement de lieux. Cette incertitude a été apposée par l'une des règles de l'outil sur la base de l'étude du contexte d'apparition de ces entités. Il importerait en conséquence de lever ce caractère d'incertitude pour *Versailles* tandis qu'il devrait être renforcé pour *mars* lorsqu'il s'agit bien du mois.

(15) Il a bénéficié d'une hospitalisation au CH de lieu_passur Versailles en lieu_passur mars où l'évolution a été favorable

Gérer l'ordre des mots. L'ordre des mots différences syntaxiques entre l'anglais et le français implique de modifier en profondeur les expressions régulières.

Ainsi, pour les noms d'hôpitaux, le schéma <nom d'hôpital + déclencheur> présent en anglais *(Washington Hospital)* est appliqué en français sur la portion « *L'examen* » en considérant le déclencheur *clinique* (exemple 16).

(16) hopital L'examen clinique retrouve bien évidemment les signes de cette insuffisance cardiaque globale.

Pour les adresses postales, le schéma <nom de voirie + déclencheur> présent en anglais *(Downing Street)* est appliqué sur la portion « *tiques par voie* » avec le déclencheur *voie* (exemple 17) ; la gestion des caractères accentués est détaillée ci-après.

(17) l'évolution a été favorable sous de fortes doses de diuré adresse_postale tiques par voie IV.

Problème de gestion de l'encodage des caractères. Ce problème *a priori* trivial est resté en suspens. Il produit ainsi deux types d'erreurs : (i) une absence d'anonymisation des informations comportant des caractères accentués, et (ii) des déclenchements étranges de certaines règles en excluant tout caractère accentué de la portion traitée (exemples 17 et 18).

6. Il existe deux villages « *Mars* » sans autre complément de lieu, dans l'Ardèche et dans le Gard.

(18) L'évolution a été marquée par la survenue d'un épisode de fibrillation atriale motivant alors un traitement par Cordarone arrêté en raison d'une [nom-famille hypothyro][lieu_passur i die].

Problème de couverture des listes. Un problème de couverture des listes d'entités nommées déclenche quelques sur-anonymisations qu'il est possible de résoudre par un nettoyage approfondi des listes. Les listes de noms et de prénoms utilisées étant internationales, elles intègrent des prénoms inutilisés en français mais dont la forme correspond à un mot de la langue (exemple 19 avec le prénom *Cher*).

(19) Mon [prénom Cher] [nom-famille Pascal],

5.3.6 Conclusion

L'adaptation de De-ID au français a concerné trois catégories d'éléments, avec un travail d'adaptation plus ou moins complexe selon la catégorie traitée.

- La recherche d'un dictionnaire de langue pour le français *(dictionnaire de noms communs)* et de listes d'entités nommées adaptées au contexte culturel français *(noms, prénoms, villes, hôpitaux, etc.)* n'a été d'aucune difficulté.
- La traduction et l'adaptation des listes de déclencheurs utilisées par De-ID n'a guère posé de problème particulier.
- En revanche, la compréhension puis l'adaptation des règles utilisées dans le programme s'est révélée particulièrement complexe, tant pour lire les expressions régulières que pour comprendre l'objectif visé par chacune de ces règles.

Compte-tenu de ces dernières difficultés et des premiers résultats obtenus, nous avons arrêté l'adaptation de De-ID au français. Nous avons estimé que poursuivre l'adaptation nous demanderait autant de temps que de produire un nouvel outil. En conséquence, nous avons décidé de développer notre propre anonymiseur, en nous inspirant des caractéristiques et des principes techniques mis en œuvre dans De-ID.

5.4 L'outil « Medina » : anonymisation nominative et numérique

5.4.1 Présentation générale

Compte tenu de l'expérience acquise lors de la tentative d'adaptation de De-ID, nous avons retenu les mêmes principes de fonctionnement pour développer notre anonymiseur. Notre outil, nommé Medina *(MEDical INformation Anonymization)*, repose intégralement sur des méthodes symboliques.

Notre outil se compose d'un script principal, de deux scripts secondaires et d'un script de post-traitement pour anonymiser un corpus (voir figure 5.1).

Étape principale : étiquetage des entités à anonymiser

Le script principal consiste à repérer dans chaque document les entités devant faire l'objet d'une anonymisation. Ce repérage repose sur l'utilisation conjointe de listes, de déclencheurs et de règles. À l'issue de ce premier traitement, le contenu de chaque document sera étiqueté au moyen de balises XML encadrant les entités identifiées avec indication du type d'entité.

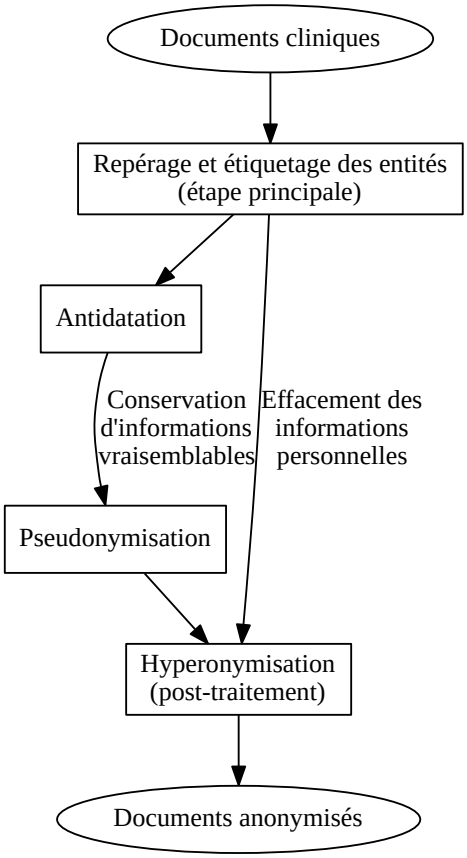


FIGURE 5.1 – Architecture globale de l'outil Medina

Étapes secondaires : antidatation et pseudonymisation

Les scripts secondaires relèvent d'un usage facultatif et permettent de gérer le type de représentation voulue des documents anonymisés. En effet, si l'objectif poursuivi consiste à masquer toute information personnelle *(remplacer les informations par des balises XML génériques)*, l'usage de ces scripts est inutile. Ces scripts ont été conçus pour conserver une représentation de sortie lisible par un humain, tout en préservant la vie privée du patient initial. Ils permettent ainsi (branche de gauche de la figure 5.1) :

1. d'antidater toutes les mentions temporelles d'un document, en conservant l'écart temporel entre deux dates d'un même document ;

2. de remplacer toutes les occurrences de noms et de prénoms par des pseudonymes *(parmi les dix noms et prénoms les plus portés en France)*, tout en assurant que chaque pseudonyme est utilisé pour la même occurrence à l'intérieur d'un document ;

Étape de post-traitement : hyperonymisation

Le script de post-traitement permet de remplacer toutes les informations étiquetées lors de l'étape principale par un hyperonyme (branche de droite de la figure 5.1). Ainsi, la portion étiquetée « *le patient habite <ville>Versailles</ville>* » devient « *le patient habite <ville />* ». Pour le cas où les scripts secondaires auraient été utilisés, les informations produites par ces scripts sont conservées (branche de gauche de la figure 5.1).

5.4.2 Ressources utilisées

Réutilisation de ressources existantes

Puisque nous avons conçu Medina comme relevant des méthodes symboliques d'une part, et que son fonctionnement s'inspire de celui de De-ID d'autre part, une grande part des ressources nécessaires à Medina a pu être reprise des expériences d'adaptation précédentes. Toutes les ressources linguistiques extérieures au programme De-ID ont ainsi pu faire l'objet d'une réutilisation directe par Medina. Ces ressources concernent les listes d'entités nommées *(en particulier pour les noms, prénoms, villes, et noms d'hôpitaux)*, le dictionnaire de noms communs du français, et la liste de déclencheurs. Nous nous sommes inspirés des distinctions faites entre listes ambigües et non ambigües de l'outil De-ID pour nettoyer ces différentes listes. Nous avons ainsi éliminé tout mot ambigu *(c.-à-d. un homographe relevant d'une liste d'entités nommées et du dictionnaire de noms communs)*, de manière à disposer de listes propres. Nous déléguons aux règles du système le processus de gestion des mots ambigus. Ces ressources ont par ailleurs été étendues pour tenir compte des caractéristiques linguistiques du corpus traité, sur la base d'une étude en corpus.

Les listes. Le dictionnaire de noms communs et les listes d'entités nommées servent dans deux cas de figure bien distincts.

Listes d'entités nommées. En ce qui concerne les listes d'entités nommées, elles ont été constituées pour rassembler une majorité de noms propres devant être anonymisés. Le contenu de ces listes est alors directement projeté sur le texte pour identifier les termes à traiter. Nous avons ainsi produit six listes d'entités nommées relevant de trois principaux domaines :

- Géographie : 30 733 noms de villes françaises et 247 noms de pays, à partir de ressources trouvées sur Internet *(INSEE, Wikipédia, etc.)* ;
- Patients : 12 826 noms de famille et 23 077 prénoms, après avoir éliminé les noms et prénoms inutilisés en France et sources d'ambiguïtés *(Agace, Dragon, Masculine, Travers, etc.)*. La présence de ce type de noms et prénoms dans les listes tient au fait qu'il s'agit de listes internationales contenant une majorité de prénoms anglo-saxons. Ces listes sont disponibles sur le site de l'ABU [7] ;
- Santé : 1 996 noms d'hôpitaux rassemblés depuis un annuaire des services de soins français, [8] 109 noms de médecins identifiés en corpus et 10 870 noms de médicaments et substances actives issus du Vidal. [9]

7. http://abu.cnam.fr/DICO/, Association des Bibliophiles Universels, CNAM Conservatoire National des Arts et Métiers, Paris.

8. http://www.sanitaire-social.com/, Annuaire Sanitaire et Social.

9. http://www.vidal.fr/

Dictionnaire de noms communs. Le dictionnaire de noms communs intègre plus de 251 000 formes fléchies différentes.[10] Son utilisation ne repose pas sur une projection directe sur le document, auquel cas la majorités des termes serait relevée, mais de manière combinée aux listes. Si un terme identifié dans le document est à la fois absent des listes d'entités nommées et du dictionnaire, alors ce terme aura de fortes chances de constituer un candidat à l'anonymisation. Un prénom original ou avec une orthographe différente de celle renseignée dans les listes d'entités nommées correspond à ce cas de figure et fera l'objet d'un tel traitement.

Liste noire. Enfin, une liste noire composée de 1 220 noms propres a été créée pour rassembler les termes qui, bien qu'étant des noms propres, ne doivent pas être anonymisés de manière systématique. Cela concerne les noms de maladies *(Creutzfeldt-Jakob, Klinefelter, Parkinson, etc.)* et de procédures chirurgicales *(Joël Cohen, etc.)*.

Les déclencheurs. Les déclencheurs ont été définis pour couvrir plusieurs catégories d'information. Ils servent d'indices pour repérer des entités à anonymiser dans leur contexte gauche ou droit d'utilisation.

- Dates : nous avons listé les noms des douze mois et des sept jours de la semaine, sous forme accentuée et désaccentuée, avec ou sans majuscule à l'initiale ;
- Mesures : seize unités de mesure ont été rassemblées *(bpm, cm, g/l, kg, etc.)* ;
- Noms de personnes :
  - nous avons retenu les titres *(Docteur, docteur, Dr., Pr, etc.)* et les termes de civilité *(Monsieur, Madame, Mme, Melle, etc.)* comme déclencheur de nom de personne, en tenant compte des formes pleines et abrégées, avec ou sans majuscule à l'initiale, en étendant cette liste aux emplois erronés de certaines abréviations *(Mr, Ms, Me, etc.)*, soit une liste de 25 déclencheurs ;
  - nous avons également utilisé neuf versions de noms de métiers *(aide, aide-opératoire, anesthésiste, opérateur, etc.)* et deux grades *(interne, externe)* comme déclencheurs de noms de personnes.
- Noms de centres de soins : nous avons listé 84 termes utilisés pour désigner un centre de soins en rassemblant aussi bien les termes désignant la structure globale *(CHR, CHU, Centre Médico-Chirurgical, Centre de Rééductation et de Réadaptation Fonctionnelle, Foyer Départemental, HAD, Maternité, Pôle Santé, Thermes, etc.)* qu'un élément de cette structure *(salle, service, unité)*, en nous inspirant des noms présents dans différents annuaires ;
- Numéros : nous avons rassemblé douze versions de déclencheurs de numéros de dossier ou de référence numérique à une opération chirurgicale *(Are K, Réf., CM/Sto, dossier, etc.)*.

Production de nouvelles ressources

Les règles constituent le dernier type de ressource utilisé. Les règles utilisées dans De-ID dont nous avions commencé l'adaptation aux spécificités de la langue française et du corpus utilisé n'ont pu faire l'objet d'une réutilisation. Les règles utilisées dans notre outil Medina ont donc fait l'objet d'un développement intégral. En

10. Les formes fléchies d'un nom sont les formes de ce nom au singulier et au pluriel ; les formes fléchies d'un adjectif sont les formes de cet adjectif au masculin et au féminin, elles-mêmes au singulier et au pluriel ; enfin, les formes fléchies d'un verbe sont les différentes formes conjuguées de ce verbe.

dehors de la quasi impossibilité de recopier tout ou partie des expressions régulières implémentées dans De-ID, ce développement est justifié par deux éléments. En premier lieu, ce redéveloppement permet de traiter toutes les catégories d'informations et seulement les catégories définies par la procédure. En second lieu, produire nos propres règles nous permet de traiter efficacement les différentes catégories d'informations en nous fondant notamment sur une étude du corpus.

5.4.3 Architecture de l'étape principale

Processus de repérage et d'étiquetage

Le processus de repérage et d'étiquetage des informations à anonymiser est appliqué deux fois de suite sur chaque document, la seconde passe permettant de se focaliser plus spécifiquement sur le voisinage des informations déjà anonymisées lors de la première passe, autorisant ainsi le raffinement des opérations déjà effectuées *(un nom de famille reconnu comme tel lors de la première passe et précédé d'une initiale verra cette initiale traitée comme un prénom lors de la seconde passe).* À l'issue de ce traitement en deux passes, on obtient un document dans lequel les entités reconnues comme devant faire l'objet d'une anonymisation seront encadrées de balises XML typantes. Les scripts secondaires et de post-traitement (voir section 5.4.1) sont alors appliqués sur la sortie de ce processus.

Dans le cadre de la première passe d'anonymisation, le traitement se fait ligne par ligne, [11] en trois étapes successives (voir figure 5.2) :

1. Détection des entités numériques et assimilées *(âges, adresses postales, téléphones, dates, numéros de série, numéros de sécurité sociale, codes postaux, mesures)* par l'application de règles fondées sur l'étude du contexte des éléments numériques ;
2. Traitement de chaque token [12] *(autre qu'une balise XML, qu'un signe de ponctuation et qui soit de taille strictement supérieure à trois caractères)* par la recherche du token à l'identique dans l'une des listes d'entités nommées *(médicaments, noms, prénoms, villes)* ;
3. Application des règles pour toutes les catégories d'information *(noms, hôpitaux et unités, villes, prénoms, médicaments)*, fondée sur l'utilisation des déclencheurs et l'étude du contexte local des entités potentielles.

La seconde passe d'anonymisation travaille de nouveau sur l'intégralité du document, ligne par ligne, au moyen d'une seule étape. Cette étape repose sur l'utilisation de règles et consiste à étudier le voisinage des éléments qui ont été étiquetés comme devant faire l'objet d'une anonymisation lors de la première passe. Sur le plan qualitatif, cette seconde passe permet d'affiner le repérage et l'étiquetage des éléments à anonymiser. Sur le plan technique, elle permet de résoudre des problèmes trop complexes à représenter en une seule expression régulière. Enfin, c'est également un moyen de

11. Le choix de ce traitement repose sur le format des fichiers que nous avons traités : aucune phrase ne se trouve segmentée sur plusieurs lignes. Nous n'avons donc pas eu à gérer les entités à cheval sur deux lignes. Dans un tel cas de figure, une étape préalable de rétablissement de chaque phrase sur une seule ligne se révèle indispensable pour faciliter la tâche d'anonymisation par les méthodes symboliques.

12. Nous avons considéré comme token toute suite de caractères comprise entre deux espaces. Ainsi, les prénoms composés *(Jean-Pierre)* et les noms de ville *(Fontenay-le-Fleury)* seront considérés comme un seul token et cherchés à l'identique dans les listes.

pondérer les règles : celles de la seconde passe, parce qu'elles sont appliquées après les autres, ont un poids plus faible dans le processus global du système.

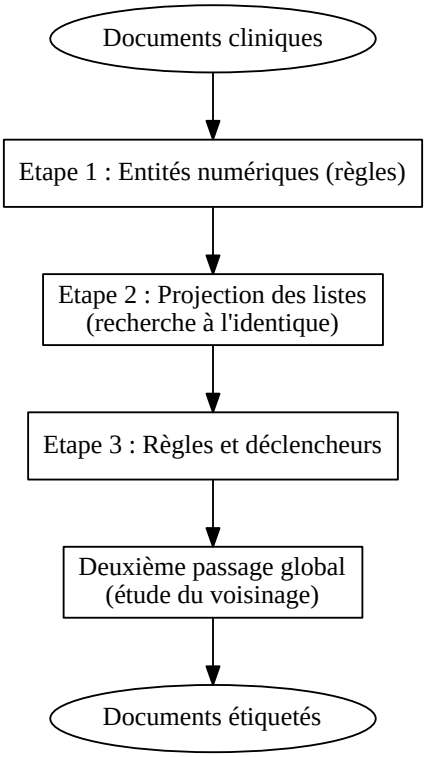


FIGURE 5.2 – Architecture détaillée de l'étape de repérage et d'étiquetage des entités

Nous renseignons dans le tableau 5.3 le nombre de règles implémentées pour chaque type d'entité, en fonction de l'étape concernée dans le traitement de la première passe d'anonymisation.

Prioritisation des règles et des listes

Au niveau de chaque ligne du document, l'application de certaines règles et listes a été définie selon un ordre précis, de manière à optimiser les anonymisations réalisées.

Ordre des règles. La prioritisation de certaines règles a été définie empiriquement comme suit :

- les entités numériques *(âges, codes postaux, dates, numéros de sécurité sociale, téléphones, etc.)* sont traitées en premier lieu, dans la perspective de se fonder sur ces premières anonymisations pour étudier le contexte *(par exemple, une entité étiquetée comme un code postal aura de fortes chances d'être suivie d'une entité relevant de la classe des noms de villes)* ;
- les entités nommées traditionnelles telles que représentées par les six listes d'entités précédemment décrites sont ensuite traitées.

	Étape 1	*Étape 2*	*Étape 3*
adresses postales	1		
âges	2		
codes postaux	1		
dates	29		
mesures	4		
noms de famille		1	5
noms de médicament		6	1
noms d'hôpitaux et d'unités			16
numéros de sécurité sociale	1		
numéros de série	4		
numéros de téléphones	2		
prénoms		1	1
villes		3	2

TABLE 5.3 – Nombre de règles par type d'entité d'après l'étape de traitement

Ordre des listes. L'application des différentes listes fait également l'objet d'un ordonnancement particulier :

- au niveau des catégories d'informations, nous appliquons en tout premier lieu les règles permettant de déterminer les noms d'hôpitaux ou de services hospitaliers avant toutes les autres, de manière à assurer la priorité des premières sur les secondes. Cela concerne les noms d'hôpitaux qui intègrent le nom de la ville d'implantation *(Centre Hospitalier de Rennes, Hôpital de Nantes, etc.)* et les noms de services hospitaliers nommés en hommage à des médecins ou des personnalités du monde médical *(Salle J. B. Bouillaud, Unité P. Soulié)* ;
- pour une même catégorie d'information, nous appliquons de manière prioritaire les règles les plus sûres *(celles reposant sur le contenu des listes d'entités nommées et/ou mobilisant les déclencheurs)* avant d'appliquer les règles potentiellement génératrices de moins bons résultats *(pour les noms de personnes, l'initiale d'un prénom suivie d'un nom, sans que ce nom ne soit présent dans la liste des noms, ni qu'un déclencheur de personne ne précède l'initiale du prénom).*

5.5 Synthèse

Dans ce chapitre, nous avons présenté trois expériences de production d'un outil d'anonymisation automatique des comptes rendus cliniques pour le français. Les informations traitées par ces outils sont de deux types : numériques *(codes postaux, dates, téléphones, numéros de série, codage des actes médicaux)* et nominatives *(noms, prénoms, villes, adresses postales, noms d'hôpitaux, marques des défibrillateurs cardiaques).* [13]

La première expérience, réalisée en 2002 à l'occasion d'un stage de fin d'études, concerne trois types de documents *(certificats médicaux initiaux, comptes rendus opé-*

13. L'anonymisation du codage des actes médicaux n'a été réalisée que sur le corpus de stomatologie, celle des marques de défibrillateurs cardiaques n'a porté que sur le corpus de cardiologie.

ratoires et lettre de correspondance) en stomatologie chirurgie maxillo-faciale. Principalement fondé sur la projection de lexiques spécialisés et l'utilisation de règles, et accompagné d'une mémoire des anonymisations réalisées pour une seconde passe d'anonymisation, le système ainsi développé a permis l'obtention de bons résultats *(rappel : 0,8298 ; précision : 0,8667 ; F-mesure : 0,8478)* sur un corpus réduit *(dix documents, 141 entités)*.

La deuxième expérience se rapporte à l'adaptation au français d'un système existant, De-ID. Si la recherche de ressources pour le français *(listes d'entités nommées et dictionnaire de noms communs)* et la traduction des déclencheurs n'a pas posé le moindre problème, la difficulté de compréhension des règles implémentées sous la forme d'expressions régulières n'a pas rendu possible l'adaptation complète au français. L'analyse des erreurs a mis en évidence des problèmes restants relatifs à l'ordre des mots, à la gestion des encodages de caractères et à la couverture incomplète des listes utilisées.

La dernière expérience concerne le développement d'un nouvel outil d'anonymisation, intitulé Medina, fondé sur les mêmes caractéristiques que celles utilisées pour développer l'outil De-ID. Notre système repose sur une première passe d'anonymisation fondée sur la projection de lexiques et l'utilisation de règles combinées à des déclencheurs, et une seconde passe d'étude du voisinage des informations déjà anonymisées dans le but d'affiner les résultats produits. L'évaluation réalisée sur un corpus de lettres de correspondance en cardiologie *(62 documents, 654 entités)* a permis l'obtention de résultats *(rappel : 0,8303 ; précision : 0,8551 ; F-mesure : 0,8425)* similaires à ceux de la première expérience, malgré les différences d'implémentation et de domaines médicaux.

Au terme de ces différentes expériences, nous avons pu constater que les méthodes symboliques permettent l'obtention de résultats de qualité mais qu'ils sont difficilement généralisables, ce qui se traduit par une précision meilleure que le rappel. Les méthodes symboliques se révèlent particulièrement adaptées pour traiter les informations numériques. Si elles permettent également de traiter correctement les informations nominatives, nous n'avons pas été en mesure de produire des règles efficaces pour deux catégories *(les adresses postales et les noms d'hôpitaux)*. Nous estimons que les approches à base d'apprentissage statistique devraient permettre de pallier ce problème.

Chapitre 6

Méthodes par apprentissage statistique

« [...] ce que j'ai entendu là, c'était pas d'la musique. C'était du bruit, rien d'autre. Enfin, vous vouliez faire quoi ?
— Un pont à partir d'une gamme de ***mi*** *pentatonique qui utilise la septième majeure comme note de passage », répondit le doyen. L'archichancelier examina la page ouverte.*
« Mais je lis ici ***Première leçon : Au clair de la lune****, dit-il.*
— Hum, hum, hum, j'étais un peu impatient », confessa le doyen.

Accros du roc
TERRY PRATCHETT

Sommaire

6.1 Introduction

Dans ce chapitre, nous présentons les différentes expériences d'anonymisation automatique de documents cliniques que nous avons réalisées au moyen d'algorithmes reposant sur un apprentissage statistique. Nous introduisons le protocole expérimental que nous avons mis en place (section 6.2), présentons le formalisme utilisé et l'outil retenu (section 6.3), puis détaillons les modalités de production des « caractéristiques » et la mise en place des expériences que nous avons menées (section 6.4). L'évaluation des résultats produits est présentée dans le chapitre suivant.

6.2 Protocole expérimental

Pour nos expérimentations à base d'apprentissage statistique, nous avons mis en place un protocole expérimental qui repose sur le principe de la validation croisée sur dix parties. [1]

La mise en place de ce protocole par validation croisée est justifiée par deux raisons. En premier lieu, cette procédure nous permet de produire un modèle plus robuste puisque produit sur un plus grand nombre de fichiers. D'autre part, puisque les résultats produits sont également évalués sur un plus grand nombre de fichiers, il nous est alors possible d'accorder plus de crédits aux évaluations réalisées.

Nous détaillons dans cette section les modalités de découpage des corpus et de répartition des parties.

6.2.1 Découpage des corpus

Afin d'autoriser des expérimentations par validation croisée, nous avons créé dix parties stratifiées issues des sous-corpus d'apprentissage et de développement, en laissant de côté le corpus de test pour les besoins de la comparaison des deux types de méthodes. Chacune des dix parties intègre respectivement 25 fichiers du corpus d'apprentissage et 25 fichiers du corpus de développement (voir tableau 6.1).

Jeu d'extraction	Premier		Deuxième	Troisième
Nombre de fichiers	100		212	250
Type d'annotation	Double		Simple	Simple
Nombre de fichiers	62	38	212	250
Sous-corpus d'affectation	Test	Apprentissage		Développement
Nombre de fichiers	62	250		250
Validation croisée en dix parties	—	25 par partie		25 par partie

TABLE 6.1 – Modalités de découpage des différents sous-corpus pour la validation croisée en dix parties

1. La validation croisée est une technique d'échantillonnage de corpus qui consiste à segmenter un corpus en n parties puis à utiliser une des n parties pour l'évaluation tandis que les $n-1$ parties restantes sont utilisées pour l'apprentissage. L'opération est répétée en utilisant pour l'évaluation une partie qui n'aura pas déjà été utilisée comme évaluation. Au final, les opérations d'apprentissage et d'évaluation seront réalisées n fois. Dans nos expériences, nous nous sommes appuyés sur une segmentation du corpus en dix parties.

6.2.2 Répartition des parties

Puisque nous disposons de dix parties et qu'il est possible de fournir aux outils CRF un sous-corpus de développement pour l'optimisation du modèle, [2] nous lançons la validation croisée à chaque tour sur la base de la répartition suivante :

- la construction du modèle se fait sur 8 parties,
- l'optimisation de la construction se fait sur la neuvième partie (développement),
- et l'évaluation par application du modèle est réalisée sur la dernière partie (test).

Chacune des dix parties se retrouve utilisée comme corpus de développement et comme corpus de test une fois dans tout le processus.

Tour	Sous-corpus		
	Apprentissage	Développement	Test
0	2, 3, 4, 5, 6, 7, 8, 9	1	0
1	0, 3, 4, 5, 6, 7, 8, 9	2	1
2	0, 1, 4, 5, 6, 7, 8, 9	3	2
3	0, 1, 2, 5, 6, 7, 8, 9	4	3
4	0, 1, 2, 3, 6, 7, 8, 9	5	4
5	0, 1, 2, 3, 4, 7, 8, 9	6	5
6	0, 1, 2, 3, 4, 5, 8, 9	7	6
7	0, 1, 2, 3, 4, 5, 6, 9	8	7
8	0, 1, 2, 3, 4, 5, 6, 7	9	8
9	1, 2, 3, 4, 5, 6, 7, 8	0	9

TABLE 6.2 – Répartition de chaque partie dans les trois sous-corpus à chaque tour

Nous résumons dans le tableau 6.2 la répartition des différentes parties dans les trois sous-corpus de la validation croisée à chaque tour. Il apparaît clairement que pour chacun des dix tours de la validation croisée, cette dernière s'effectue sur chacune des dix parties créées au préalable, chaque partie étant utilisée comme corpus d'apprentissage, de développement ou de test, tout au long du processus.

6.3 Configurations

Nos expériences d'anonymisation automatique au moyen d'algorithmes d'apprentissage statistique reposent exclusivement sur le formalisme des champs aléatoires conditionnels (CRF) de chaîne linéaire. Comme présenté en section 2.3.2, ce formalisme présente l'avantage de prendre en compte le contexte pour étiqueter une séquence de tokens, ce qui se révèle indispensable en matière d'anonymisation, ou plus généralement, dans les tâches de repérage d'entités nommées.

6.3.1 Choix de l'outil

Plusieurs algorithmes reposant sur les CRFs sont librement disponibles et utilisables parmi lesquels Mallet [McCallum, 2002], CRF++ [Kudo et al., 2004], et MIST

2. Le corpus de développement est utilisé lors de la construction du modèle pour mettre au point le modèle. Cette mise au point se fonde sur l'évaluation de l'application du modèle en cours de construction.

[Aberdeen et al., 2010]. Plus récemment, une implémentation a été réalisée au LIMSI sous le nom de Wapiti [3] [Lavergne et al., 2010].

Nous avons sommairement testé trois outils *(CRF++, MIST, et Wapiti)* et restreint la découverte approfondie du formalisme à deux outils *(CRF++ et Wapiti)*. Le choix de ces deux outils a été dicté par leur simplicité d'utilisation d'une part, et parce qu'ils partagent le format des fichiers de configuration et celui des fichiers à traiter d'autre part, réduisant ainsi le travail de préparation des données et facilitant la comparaison d'utilisation des outils.

Après une prise en main de ces deux outils et la réalisation de différents tests d'utilisation, nous avons décidé de ne retenir que l'outil Wapiti. La rapidité d'exécution [4] et la possibilité de mieux paramétrer l'outil en termes de choix de la méthode d'optimisation implémentée, de choix des valeurs du paramètre de régularisation *(ℓ^1, ℓ^2)* et de possibilité — au décodage — de recourir à des probabilités, ont milité en faveur de cet outil. La présence quotidienne de son concepteur dans les locaux du laboratoire constitue également un avantage appréciable.

6.3.2 Paramètres

Méthodes d'optimisation

L'outil Wapiti implémente plusieurs algorithmes d'optimisation pour la recherche des paramètres du modèle construit *(L-BFGS, OWL-QN, SGD-L1, BCD et RPROP)*. [5] Parmi les algorithmes disponibles, nous avons retenu RPROP [Riedmiller et Braun, 1993]. Ce dernier repose sur le principe des réseaux neuronaux. Il est notamment reconnu pour sa rapidité de fonctionnement.

Réglage des pénalités ℓ^1 et ℓ^2

La pénalité ℓ^1 *(priorité laplacienne, utilisée comme paramètre de sélection des caractéristiques)* étant réglée par défaut à 0,5, nous avons testé les différentes valeurs de cette régularisation dans l'intervalle [0,1;1,5]. Nous avons appliqué cette variation sur l'expérience d'anonymisation donnant lieu aux meilleurs résultats. Dans cette optique, et partant d'une régularisation ℓ^1 fixée à 0,5, nous avons cherché à trouver la valeur qui permettrait d'améliorer nos résultats. Les résultats de ces différents tests nous ont conduit à retenir la valeur ℓ^1 de 0,1 (soit la valeur la plus basse), ce qui offre à l'algorithme les marges de manœuvre les plus importantes et notamment celle de faire un sur-apprentissage. Nous n'avons pas modifié la valeur par défaut de la pénalité ℓ^2 *(priorité gaussienne, utilisée pour assurer au modèle une bonne stabilité et réduire la sur-adaptation)*.

3. http://wapiti.limsi.fr/

4. Notamment au moyen d'une option de paramétrage permettant du multi-thread à la construction du modèle. Une expérience de construction d'un modèle reposant sur l'utilisation des mêmes caractéristiques a demandé 59 min à CRF++ contre 3 min 30 à Wapiti sans multi-thread et environ 1 min 25 en utilisant les quatre cœurs du serveur sur lequel nous avons réalisé nos expérimentations.

5. Deux méthodes d'optimisation non linéaires quasi-Newton : (i) L-BFGS : limited-memory BFGS *(Broyden–Fletcher–Goldfarb–Shanno)* et (ii) OWL-QN : Orthant-wise limited-memory quasi-Newton, variante de L-BFGS pour adapter les modèles reposant sur la pénalité ℓ^1 ; une méthode non linéaire reposant sur les gradients, SGD-L1 : Stochastic Gradient Descent training for ℓ^1-regulized ; BCD : Blockwise Coordinate Descent ; et une méthode issue des réseaux neuronaux, RPROP : Resilient Propagation.

Corpus d'entraînement et de développement

La construction d'un modèle statistique reposant sur un corpus d'entraînement et l'utilisation combinée d'un corpus de développement permettent d'optimiser les décisions prises. Alors qu'en utilisant un seul corpus, l'algorithme va évaluer le modèle qu'il construit sur le même corpus que celui à partir duquel il les construit, l'utilisation d'un corpus de développement permet de disposer d'un second corpus pour l'évaluation. L'utilisation de ces deux corpus permet ainsi d'assurer une meilleure robustesse des règles construites puisqu'elles auront été évaluées sur un autre jeu de données que celui utilisé pour les produire.

Probabilités au décodage

Enfin, l'outil Wapiti permet de recourir à l'utilisation de probabilités lors de l'application du modèle sur les données de test. [6] L'utilisation de cette option sur nos données de test n'a cependant pas permis d'améliorer les résultats. Parce qu'elle avait tendance a dégrader globalement nos résultats, nous ne l'avons pas utilisée.

6.4 Expérimentations

Dans cette section, nous détaillons de quelle manière nous avons produit les caractéristiques relatives à chaque token pour permettre la construction du modèle par apprentissage. Nous présentons également les différentes conditions d'utilisation de ces caractéristiques que nous avons mises en place dans chacune des expériences.

Le modèle que nous voulons construire doit être capable de prédire dix classes, [7] plus une classe pour les tokens qui ne relèvent pas de l'une de ces dix classes. L'encodage de cette information dans le modèle *(avec une étiquette par token)* repose sur le schéma d'annotation BIO (voir section 2.3.3), ce qui impose deux étiquettes par classe *(B-classe et I-classe)* [8] plus l'étiquette O pour les tokens ne relevant d'aucune classe, soit un total de 20 étiquettes. Les modèles construits dans nos expériences reposent sur 168 149 blocs et 28 308 580 caractéristiques.

6.4.1 Caractéristiques produites sur chaque token

Dans cette section, nous détaillons les caractéristiques que nous avons fournies à Wapiti au travers du fichier de configuration (voir annexe D) pour construire le modèle. Nous distinguons différents types de caractéristiques (figure 6.1).

Individus

Nous considérons comme « token » toute suite de caractères comprise entre deux blancs typographiques. Du fait du type de tokénisation effectuée — l'ajout d'une espace typographique autour de chaque signe de ponctuation —, un token se composera, soit de caractères alphanumériques, soit de signes de ponctuation isolés. Nous retenons (voir tableau 6.3) :

6. L'option -p utilisée au décodage met en place cette utilisation.

7. Les dix classes à prédire sont les suivantes : *adresse, codepostal, date, hopital, info, nom, numero, prenom, telephone, ville*.

8. Sauf pour la classe *codepostal*, toujours de taille un token, soit une seule étiquette *B-codepostal*.

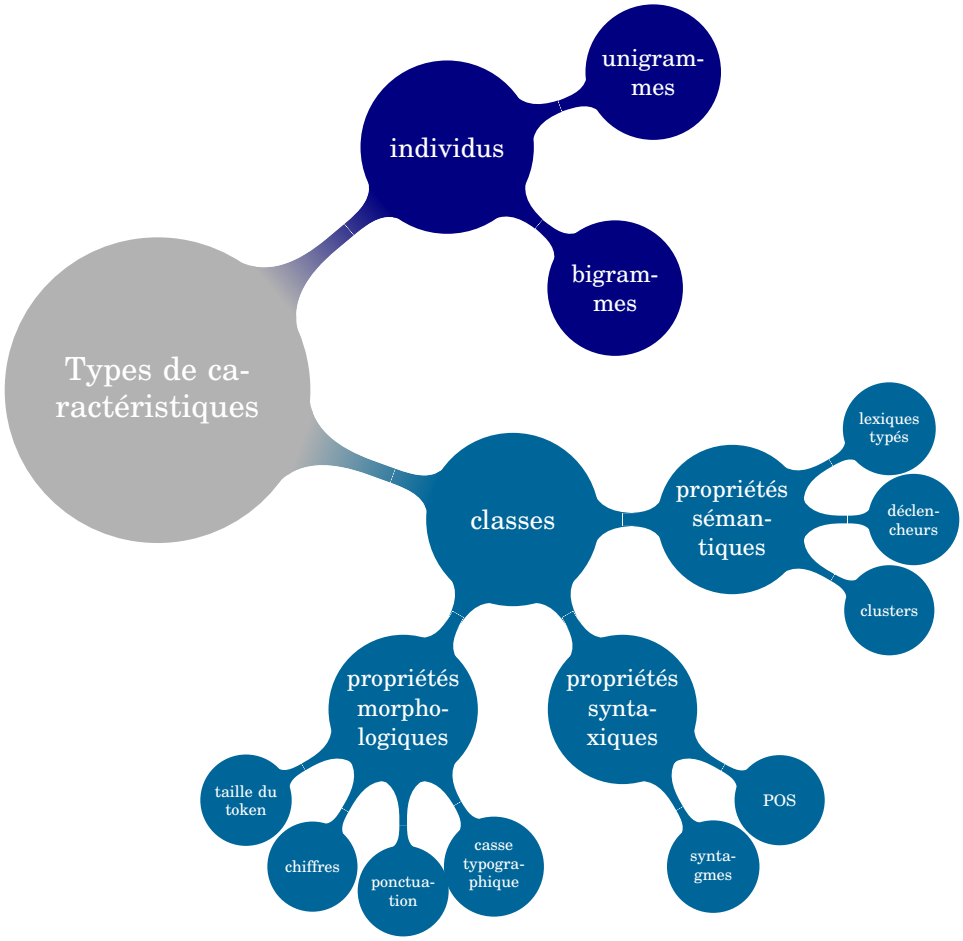


FIGURE 6.1 – Résumé des différents types de caractéristiques produites

– l'unigramme du token étudié et celui de chacun des trois tokens qui le précèdent, soit les tokens portant les numéros d'ordre suivants : « -3 », « -2 », « -1 » et « 0 » ;
– et les bigrammes de tokens autour du token étudié à raison des séquences suivantes : « -2/-1 », « -1/0 » et « 0/+1 ».

Propriétés sémantiques

Le dernier type de caractéristiques utilisées nécessite le recours à des outils ou des ressources externes. À ce titre, l'utilisation de ces informations constitue une forme d'hybridation entre méthodes symboliques et apprentissage statistique. Nous avons constitué les groupes de caractéristiques suivants pour chaque token.

Listes typées et dictionnaires. Nous vérifions la présence de chaque token dans des listes d'entités nommées et dans un dictionnaire des formes fléchies du français. Ces ressources sont les mêmes que celles utilisées par Medina dans les approches symboliques. Nous identifions ainsi : (i) la présence du token dans un lexique *(liste de noms, de prénoms, d'hôpitaux et de villes)* et (ii) l'absence du token dans un dictionnaire des formes fléchies du français.

Unigrammes	Bigrammes
votre	... / ... / votre patient
patient	... / votre patient / patient Mr
Mr	votre patient / patient Mr / Mr Platel
Platel	patient Mr / Mr Platel / Platel Alphonsie
Alphonsie	Mr Platel / Platel Alphonsie / Alphonsie né
né	Platel Alphonsie / Alphonsie né / né le
le	Alphonsie né / né le / le 20
20	né le / le 20 / 20 .
.	le 20 / 20 . / . 04
04	20 . / . 04 / 04 .
.	. 04 / 04 . / . 1953
1953	04 . / . 1953 / ...

TABLE 6.3 – N-grammes de tokens d'une phrase tokénisée

Déclencheurs. Nous vérifions que le token est : (i) un déclencheur de nom *(Docteur, Dr, Madame, Mme, M., Mrs, ME, etc.)* ou (ii) d'hôpital *(Clinique, CHU, CHR, Fondation, Hôpitaux, Polyclinique, Salle, Service, Unité, etc.)*, d'après des listes de déclencheurs produites pour l'outil Medina.

Groupes de tokens. Un dernier type de caractéristique fourni au CRF concerne des groupes *(clusters)* automatiquement constitués à partir des tokens des documents par l'implémentation de l'algorithme de clustering de Brown [Brown et al., 1992] réalisée par [Liang, 2005]. Cet algorithme cherche à optimiser les clusters pour minimiser la perplexité d'un modèle de langue construit sur les bigrammes de ces clusters. Nous avons lancé cet algorithme sur l'ensemble des tokens des documents du corpus *(apprentissage, développement et test)* en spécifiant deux paramètres : (i) la fréquence d'apparition minimale de chaque token et (ii) le nombre souhaité de classes finales. Afin d'éliminer les tokens de faible occurrence, nous avons fixé la fréquence d'apparition minimale à quatre occurrences. D'autre part, nous avons empiriquement établi le nombre de classes finales à 320, une valeur plus élevée ou plus faible conduisant à des regroupements ne faisant pas sens.

Il resulte de l'application de cet algorithme une classification des 3 163 tokens les plus fréquemment utilisés dans le corpus traité. Les 320 classes générées intègrent un nombre hétérogène de tokens *(entre 1 et 71 tokens)*. Nous donnons dans le tableau 6.4 un exemple de contenu de quelques groupes générés. Nous observons que les classes générées révèlent un rassemblement sémantique globalement exact, en particulier les quatre derniers groupes qui rassemblent respectivement les déclencheurs de noms, tous les mois de l'année *(et uniquement les mois de l'année)*, une majorité de noms et de prénoms *(et cinq noms communs)*, et les années *(y compris sur deux chiffres ou avec le chiffre zéro remplacé par la lettre "o")*, groupes que nous estimons particulièrement utiles pour une tâche d'anonymisation. En revanche, faute de disposer d'un nombre suffisant d'occurrences, les noms de villes présents dans le corpus *(Laval, Lorient, Paimpol, Rennes)* ont été regroupés dans deux groupes avec d'autres tokens, parce qu'ils partagent des contextes communs, sans qu'émerge de ces groupes un sens précis.

Sur la base des regroupements effectués, nous observons que ces groupes rassemblent donc des éléments, soit devant faire l'objet d'une anonymisation *(groupe des*

mois et des années dans le tableau 6.4), soit fournissant un indice sur le traitement à appliquer au mot qui suit *(groupe des déclencheurs de noms)*. Les groupes de tokens générés automatiquement apportent donc au CRF des indices utiles en matière de repérage d'entités partageant des contextes d'apparition et devant être anonymisées.

Identifiant numérique	Nombre de mots	Exemples de tokens appartenant au groupe généré
1000	1	l'
000010	10	anti-agrégant, antithrombine, esv, extra-systoles, extrasystoles, fréquences, grade, linéaires, stade, timi
0011100	5	angio-irm, écho-doppler, échodoppler, état, unité
00111010	1	examen
01110000	16	lorient, paimpol, rennes [...] 24h, dipyridamole, diurétiques, vésiculaire [...]
01110001	21	laval [...] cardiomégalie, complication, dépistage, galop, malaise, palpitation, phlébite, succès [...]
11100001	6	“interne,unité", madame, mademoiselle, mme, monsieur, mr
111111010	13	aout, août, avril, février, janvier, juillet, juin, mai, mars, novembre, octobre, septembre, décembre
111111100	71	anaïs, antonin, bertin, dumontier, jean-baptiste, leclercq, legrain, lison, morisseau, pierre [...] basal, enfant, ergométrique, général, hémodynamique
11111111101	16	05, 1982, 1994, 2000, 2oo5, 2005, 2005), 2006 [...]

TABLE 6.4 – Exemple de contenu de quelques groupes (clusters) générés par l'algorithme de clustering de Brown et nombre de mots différents dans ces groupes

Propriétés syntaxiques

Nous attribuons à chaque token l'étiquette en partie du discours qui lui correspond et tentons d'inférer de ces étiquettes le syntagme duquel dépend le token.

Étiquetage en parties du discours. Nous procédons dans un premier temps à un étiquetage en parties du discours d'après une simplification des étiquettes fournies par le TreeTagger [Schmid, 1994]. Nous conservons l'étiquette fournie, même si elle peut être erronée contextuellement du fait des probabilités d'association. Des solutions de correction existent,[9] notamment en appliquant le logiciel Flemm [Namer, 2000] sur les sorties du Tree Tagger.

De manière alternative, nous avons également utilisé l'étiqueteur de Brill [Brill, 1992] réentraîné pour le français sur un corpus d'articles du journal *Le Monde* [Allauzen et Bonneau-Maynard, 2008] avec le jeu d'étiquettes des parties du discours utilisées lors de la campagne d'évaluation GRACE [Adda et al., 1999]. Si la qualité de l'étiquetage effectué est manifeste, la richesse du jeu d'étiquettes employé se révèle

9. Le temps de travail nécessaire estimé pour mettre en place une chaîne de correction des sorties du Tree Tagger, au regard du faible gain que nous pensons obtenir en matière d'anonymisations qui en découleraient, nous ont conduit à ne pas nous engager, pour l'instant, dans cette voie.

cependant contre-productive dans le sens où — même simplifiée [10] — elle revient à utiliser un paradigme de labels trop important et conduit l'outil CRF, quel qu'il soit, [11] à limiter le nombre de caractéristiques utilisant ces informations.

L'utilisation des parties du discours permet de distinguer les catégories qui ne font jamais l'objet d'une anonymisation *(adjectifs, adverbes, prépositions, pronoms, verbes)*, de celles qui font l'objet d'une anonymisation quasi systématique *(noms propres)* et de catégories intermédiaires pour lesquelles le traitement dépend du contexte *(déterminants,* [12] *noms communs, numéros)*.

Inférences de syntagmes. Dans un second temps, nous estimons l'appartenance du token à un syntagme sur la base des étiquettes précédemment produites. De manière empirique et en nous inspirant des démarches entreprises dans des travaux récents sur les chunks en français [Tellier et al., 2012, Tellier, 2012], nous rassemblons sous l'appellation « syntagme nominal » les séquences de déterminants, de noms communs et propres, de nombres et de pronoms ; le « syntagme verbal » se compose des seuls verbes ; le « syntagme prépositionnel » se compose des seules prépositions.

Notre démarche est simplifiée à l'extrême car seuls nous intéressent réellement les syntagmes nominaux. Cette propriété est envisagée pour permettre (i) de restreindre l'appartenance des entités à anonymiser aux syntagmes nominaux, et (ii) de circonscrire syntaxiquement les frontières des portions à anonymiser. Nous avons ainsi pu constater sur le corpus d'apprentissage que sur les 16 748 tokens que nous rattachons à un syntagme verbal, aucun n'appartient à une catégorie d'information à anonymiser. Nous effectuons le même constat concernant les 16 762 tokens rattachés à un syntagme prépositionnel.

Pour peu que l'étiquetage d'origine du token en « verbe » et « préposition » par le Tree Tagger soit correct, [13] le rattachement d'un élément à un syntagme verbal ou à un syntagme prépositionnel paraît donc fortement discriminant et utile pour exclure un token d'une portion anonymisée. Les erreurs d'étiquetage du Tree Tagger nécessiteraient une étape de correction dans certains cas bien particuliers, de manière à éviter les erreurs de rattachement des tokens à un syntagme. Nous relevons notamment le cas de l'initiale des prénoms commençant par « A » que le Tree Tagger considère comme une préposition (désaccentuée). Nos traitements nous conduisent donc à rattacher l'initiale « A » à un syntagme prépositionnel, ce qui ne permet pas au CRF de traiter ce token *(99,9 % des tokens rattachés à un syntagme prépositionnel ne nécessitent pas une anonymisation, le 0,1 % restant relève de la préposition contractée « au » dans les intervalles de dates)*.

Propriétés morphologiques

Les caractéristiques de surface correspondent aux propriétés qu'il est possible d'inférer directement de chaque token sans recourir à des ressources ni outils ex-

10. Nous avons simplifié les étiquettes de l'action GRACE aux deux premiers caractères qui composent chaque étiquette : ainsi, « Ncms » *(nom commun masculin singulier)* est simplifiée en « Nc » et l'étiquette « Vmip3s » *(verbe principal, indicatif présent, 3e personne du singulier)* est réduite à « Vm ».

11. Comportement observé aussi bien avec CRF++ qu'avec Wapiti.

12. Le déterminant est anonymisé, par exemple lorsqu'il constitue une particule dans un nom de famille *(M. Le Cerf)* ou lorsqu'il est constitutif d'un nom de ville *(Le Mesnil)*.

13. Sur les 16 748 tokens rattachés aux syntagmes verbals, nous relevons 85 tokens (soit 0,50 %) qui le sont par erreur. Ont ainsi été étiquetés « verbe » par le Tree Tagger 24 noms de famille intégralement écrits en majuscules et initiales de prénom ou encore 14 années. La même proportion d'erreurs se retrouve pour la catégorie « préposition » *(99,5 % d'étiquetage correct pour 0,5 % d'étiquetage erroné)*.

ternes. Nous avons ainsi inféré les propriétés suivantes de chaque token :

- la casse typographique parmi trois classes *(tout en majuscules, majuscule à l'initiale puis minuscules, tout en minuscules)*. Cette propriété permet de distinguer les noms propres *(commençant généralement par une majuscule)* des autres mots. La casse relevée est celle :
  - du token étudié,
  - et des bigrammes de tokens relevant de l'une des deux séquences suivantes : « -1/0 » et « 0/+1 ».
- le token est-il un signe de ponctuation *(propriété binaire)*. Cette propriété est notamment utile pour certaines entités numériques *(numéros de téléphone et dates avec un séparateur généralement constitué d'une ponctuation)* ;
- le token est-il composé de chiffres *(propriété binaire)*. Cette propriété permet de mettre en évidence les entités numériques *(dates, numéros de téléphone)* ;
- la taille du token en nombre absolu de caractères. Cette propriété est envisagée pour distinguer les mots outils *(composés d'un nombre réduit de caractères)* des autres mots *(potentiellement des entités à anonymiser)*.

Nous donnons dans le tableau 6.5 un exemple des propriétés calculées par Wapiti selon ces directives pour chaque token. Le format de représentation des informations correspond au schéma d'annotation BIO [14] décrit en section 2.3.2.

Tokens	Casse	Ponctuation	Chiffres	Taille
votre	I-mm	O	O	5
patient	I-mm	O	O	7
Mr	B-Mm	O	O	2
Platel	I-Mm	O	O	6
Alphonsie	I-Mm	O	O	7
né	B-mm	O	O	2
le	I-mm	O	O	2
20	O	O	B-digit	2
.	O	B-punct	O	1
04	O	O	B-digit	2
.	O	B-punct	O	1
1953	O	O	B-digit	4

TABLE 6.5 – Propriétés morphologiques calculées par Wapiti pour chaque token

6.4.2 Mise en place des expérimentations

Nous nous sommes inspirés des indications fournies par [McCallum, 2003] (voir section 2.3.2) pour établir plusieurs expériences d'anonymisation reposant sur les CRF. Compte-tenu de la distinction que nous avons faite entre les différents types de caractéristiques (section 6.4), nous avons ainsi réalisé des expériences, soit avec un seul type de caractéristiques, soit en combinant plusieurs de ces types. Nous présentons dans les paragraphes suivants les motivations qui nous ont poussé à choisir les configurations de nos différentes expérimentations.

14. Pour rappel, « B » *(begin)* renvoie au début d'une séquence de caractéristiques, « I » *(in)* à l'intérieur d'une séquence et « O » *(out)* à un token hors de toute séquence.

Utilisation d'un seul type de caractéristiques

Le premier lot d'expériences consiste à restreindre la construction du modèle à un seul type de caractéristiques :

- les caractéristiques de surface uniquement *(casse typographique, composition du token en termes de chiffre ou de ponctuation, taille du token)*. Le modèle fait le lien entre ces caractéristiques et la réponse attendue ;
- l'utilisation des ressources externes uniquement. Les ressources externes renvoient aux éléments suivants : (i) l'appartenance du token à une liste *(noms, prénoms, villes, hôpitaux, dictionnaire de formes fléchies)*, (ii) l'étiquette du token en partie du discours fournie par le Tree Tagger, et (iii) l'identifiant numérique du cluster fourni par l'algorithme de Brown. Le modèle est alors construit en apprenant le lien qui existe entre la réponse attendue et l'une des trois propriétés précédemment listées ;
- les formes des tokens uniquement : le modèle est construit uniquement sur la base des tokens tels qu'ils se présentent dans les documents sans recourir à d'autres types d'information. Le modèle fait donc le lien entre la forme du token et la réponse attendue.

L'objectif de ces premières expériences vise à mettre en évidence quels résultats de base il est possible d'atteindre avec des approches simples.

Combinaison de plusieurs types de caractéristiques

Le deuxième ensemble d'expériences consiste à combiner les types de caractéristiques testés précédemment de manière isolée :

- les formes des tokens combinées aux caractéristiques de surface,
- les formes des tokens combinées aux ressources externes,
- ou la combinaison maximale constituée des trois types de caractéristiques : les formes des tokens, les caractéristiques de surface et les ressources externes.

La démarche visée par ces combinaisons consiste à présenter l'intérêt qu'il y a à utiliser des caractéristiques relevant de plusieurs types d'une part, et à constater les améliorations et dégradations apportées par ces combinaisons sur chacune des catégories d'autre part.

Création de modèles propres à chaque catégorie

Contrairement aux expérimentations précédentes dans lesquelles nous avons à chaque fois construit un modèle global, le troisième ensemble d'expériences consiste à construire autant de modèles qu'il y a de catégories d'information à traiter dans le corpus *(soit un total de neuf modèles)*. Cette création de modèles distincts repose néanmoins sur l'utilisation des mêmes caractéristiques *(formes des tokens, caractéristiques de surface et ressources externes)* et des mêmes configurations d'utilisation de ces caractéristiques que celles utilisées pour le modèle global.

Nous avons également essayé de modifier les configurations d'utilisation de ces caractéristiques en cherchant une configuration dont les résultats permettraient de dépasser ceux obtenus par la configuration dite « optimale » jusqu'à présent utilisée avec le modèle global.

L'objectif de ce type d'expérimentation vise à étudier les interactions et les confusions entre catégories. Autrement dit, dans quelle mesure le modèle global se sert

des informations et des annotations sur les autres catégories pour traiter une catégorie donnée, et inversement, dans quelle mesure l'utilisation de ces informations ne perturbe-t-elle pas le traitement de chaque catégorie.

Enchaînement de méthodes

Enfin, nous avons testé un autre type d'hybridation, celui qui consiste à enchaîner les deux approches, d'abord celle à base d'apprentissage, suivi de la méthode symbolique. Nous avons ainsi appliqué Wapiti en validation croisée en dix parties, puis nous avons lancé Medina sur les sorties produites par Wapiti *(donc, sur les cinquante fichiers de chacune des dix parties de la validation croisée).*

L'objectif de cette cascade de systèmes consiste donc à tirer parti de la combinaison des deux approches. Nous avons fait le choix d'appliquer en premier lieu les approches à base d'apprentissage, parce qu'elles permettent l'obtention de bien meilleurs résultats (section 7.3) que les méthodes symboliques. D'autre part, sachant que Medina est appliqué sur les sorties annotées produites par Wapiti, l'application d'un système à base de règles ne sera donc en mesure de traiter que les informations qui n'auront pas été gérées par l'approche par apprentissage. Autrement dit, le second système *(Medina)* ne corrige pas les erreurs du premier *(Wapiti)*, mais il complète les anonymisations réalisées [Grouin et Zweigenbaum, 2011].

6.5 Synthèse

Dans ce chapitre, nous avons présenté les différentes modalités de réalisation des expériences d'anonymisation automatique au moyen des méthodes à base d'apprentissage statistique.

Dans un premier temps, nous avons présenté notre protocole expérimental. Afin de disposer d'un modèle plus robuste d'une part, et d'accorder plus de crédit aux résultats obtenus d'autre part, nous avons mis en place une procédure de validation croisée en dix parties. Nous avons ainsi détaillé la réalisation de cette procédure.

Dans un deuxième temps, ayant fait le choix du formalisme des champs aléatoires conditionnels (CRF), nous avons présenté l'outil que nous avons utilisé qui implémente ce formalisme : Wapiti. Le choix de cet outil repose sur sa rapidité de traitement et les nombreuses possibilités de paramétrage offertes. Nous indiquons quelles sont les options de paramétrage que nous avons retenues pour nos expérimentations. Les modèles construits reposent sur 20 étiquettes (selon le schéma d'annotation BIO), 168 149 blocs et 28 308 580 caractéristiques.

Enfin, nous avons présenté les différents ensembles de caractéristiques que nous avons produits, et les différentes expérimentations que nous avons définies, sur la base de modalités d'utilisation distinctes de ces caractéristiques.

Chapitre 7

Évaluations et discussion

— De quoi vous parlez ? demanda l'archichancelier.
— Je voulais juste faire remarquer l'improbabilité intrinsèque de...
— Taisez-vous, le coupa l'archichancelier, terre-à-terre.

Le faucheur
TERRY PRATCHETT

Sommaire

7.1 Introduction

Dans ce chapitre, nous présentons les résultats obtenus par les systèmes des deux approches que nous avons testées et évaluées, l'approche symbolique d'une part, et celle à base d'apprentissage statistique d'autre part. Nous discutons également des résultats obtenus et des performances réalisées par les systèmes de ces deux approches. Les mesures d'évaluation utilisées *(rappel, précision et F-mesure)* sont définies en section 3.2.2, les mesures « globales » correspondent aux macro-moyennes (section 3.2.3) et la simulation de Monte Carlo est présentée en section 3.5.2.

Nous rappelons que le sous-ensemble *(constitué par tirage aléatoire)* utilisé du corpus en cardiologie a été scindé en trois sous-corpus (voir section 4.3.3) :

- Le corpus d'apprentissage (250 fichiers) a servi à mettre au point les règles de l'outil Medina *(méthodes symboliques)* et à construire le modèle CRF *(approche par apprentissage statistique)* ;
- Le corpus de développement (250 fichiers), constitué ultérieurement avec d'autres fichiers, n'a été utilisé que lors de la construction du modèle CRF comme élément de validation du modèle en cours de construction, pour assurer une meilleure robustesse au modèle ainsi créé ;
- Enfin, la comparaison des performances réalisées sur ces deux approches est rendue possible par le corpus de test composé de 62 fichiers.

Concernant les expériences à base d'apprentissage statistique, la validation croisée en 10 parties repose sur un découpage stratifié des sous-corpus d'apprentissage et de développement, le corpus de test demeurant intact pour la comparaison des approches.

Au fur et à mesure de l'avancement de ce travail notamment lors de l'apprentissage, nous nous sommes rendu compte que certaines catégories posaient problème. Alors que la définition d'origine des noms *(le nom de famille)* et des prénoms *(le prénom complet ou l'initiale)* semblait évidente, la réalité en corpus nous a conduit à reconsidérer cette vision, particulièrement dans le cas où les initiales du prénom et du nom d'une personne sont regroupées en un seul token que le guide d'annotation nous recommande d'annoter globalement comme un nom, quand bien même l'une des deux initiales se rapporte au prénom.

- Prénom complet et nom de famille : prénom *Pierre* nom *Martin*
- Initiale du prénom et nom de famille : prénom *P.* nom *Martin*
- Initiales groupées des prénoms et noms : nom *PM*

Sur la base de ces observations, la distinction entre les catégories *nom* et *prénom* ne nous apparaissait plus justifiée, d'autant plus que le maintien de cette distinction dans le cadre du projet dans lequel s'inscrit ce travail de thèse ne présente pas d'intérêt. En conséquence, nous avons décidé de regrouper ces deux catégories sous une seule catégorie générique, et avons procédé à deux évaluations des outils, l'une tenant compte de l'existence des deux catégories *(que nous appellons « évaluation multi-catégories »)*, l'autre fondée sur la catégorie générique uniquement *(présentée comme « évaluation après fusion nom/prénom »)*.

D'autre part, la catégorie des adresses postales présente un problème de définition. Nous avons en effet donné de cette catégorie une définition fondée sur l'interprétation humaine de ce que représente une adresse *(un bloc complet intégrant le*

numéro, le type de voirie et le nom de la voie), et non une définition structurelle reposant sur les différents éléments constitutifs de l'adresse *(un numéro, un type de voirie « rue, avenue, route », un nom qui peut être celui d'une personne « Victor Hugo », d'une ville « Versailles », un nom commun « l'école de médecine », le nom d'un lieudit, etc.)*.

– Définition humaine : adresse *15 rue Victor Hugo*

– Définition structurelle : numéro *15* type *rue* prénom *Victor* nom *Hugo*

Bien que conscient de ce problème, nous n'avons pas poussé plus en avant les expériences de redéfinition de cette catégorie, en raison du faible nombre d'adresses présentes en corpus *(à rapporter à la charge de travail que cela représente)* d'une part, et parce que nous avons principalement axé ce travail de thèse sur l'anonymisation des noms de personnes.

7.2 Méthodes symboliques

Dans cette section, nous présentons les résultats obtenus par l'outil Medina présenté en section 5.4 sur le corpus de test en cardiologie.

7.2.1 Évaluation multi-catégories

Nous donnons dans le tableau 7.1 les résultats globaux obtenus par Medina sur le corpus de test en cardiologie *(62 documents)*. Sur la base de ces résultats, nous avons également calculé l'intervalle de confiance dans lequel évolue la F-mesure, selon la simulation de Monte Carlo décrite en section 3.5.2. Nous renseignons dans le tableau 7.2 les résultats détaillés pour chacune des catégories traitées.

Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure	Intervalles de confiance
548	87	110	0,8328	0,8630	0,8476	[0,8266;0,8687]

TABLE 7.1 – Évaluation globale de Medina sur le corpus de test en cardiologie

Catégorie	Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure
Dates	208	18	30	0,874	0,920	0,897
Noms	186	20	19	0,907	0,903	0,905
Prénoms	101	29	8	0,927	0,777	0,845
Hôpitaux	16	16	27	0,372	0,500	0,427
Villes	11	5	11	0,500	0,688	0,579
Codes postaux	8	0	0	1,000	1,000	1,000
Adresses	1	2	7	0,125	0,333	0,182
Téléphones	8	0	0	1,000	1,000	1,000
Appareillage	3	2	7	0,300	0,600	0,400
Numéro de série	1	0	2	0,333	1,000	0,500

TABLE 7.2 – Évaluation détaillée de Medina sur le corpus de test en cardiologie

7.2.2 Évaluation avec fusion nom/prénom

La même évaluation a été de nouveau faite après avoir rassemblé les catégories *nom* et *prénom* sous une seule catégorie générique *nom*, dans la référence et dans le fichier de sortie de l'outil Medina. Ce rassemblement de catégories ne porte donc que sur les résultats finaux.

L'outil ayant été conçu pour distinguer les noms des prénoms *(notamment au moyen de règles contextuelles reposant sur l'existence de ces deux catégories)*, nous n'avons pas modifié l'outil de telle sorte qu'il rassemble directement tous les noms et prénoms dans cette nouvelle classe générique *nom*.

Nous donnons dans le tableau 7.3 les résultats globaux après avoir rassemblé ces deux catégories, et dans le tableau 7.4, le détail des résultats obtenus sur cette catégorie générique *nom (les résultats des autres catégories ne changent pas puisque l'outil reste le même et que seule la sortie finale a été modifiée)*.

Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure	Intervalles de confiance
469	61	91	0,8375	0,8849	0,8606	[0,8386;0,8825]

TABLE 7.3 – Évaluation globale de Medina sur le corpus de test en cardiologie après avoir rassemblé les catégories nom et prénom

Catégorie	Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure
Noms	208	23	8	0,963	0,900	0,931

TABLE 7.4 – Évaluation détaillée de Medina sur le corpus de test en cardiologie après avoir rassemblé les catégories nom et prénom

7.3 Méthodes par apprentissage statistique

Afin de permettre une stricte comparaison des résultats produits par Medina et par Wapiti, nous fournissons l'évaluation du modèle Wapiti construit sur le corpus d'apprentissage *(250 fichiers)* avec l'aide du corpus de développement *(250 fichiers)* appliqué sur le corpus de test *(62 fichiers)*. À l'instar des évaluations effectuées sur les méthodes symboliques, nous avons réalisé deux évaluations, la première reposant sur l'ensemble des catégories, la seconde après fusion des catégories *nom* et *prénom*.

Nous présentons également les résultats de la validation croisée en 10 parties. Cette procédure de validation croisée permet d'assurer au modèle ainsi construit une meilleure robustesse. Précisons que cette évaluation permet également de généraliser les performances du modèle dans la mesure où l'évaluation porte sur 500 documents et non plus seulement 62. L'évaluation portant sur 10 tours, nous sommes alors en mesure d'obtenir des intervalles de confiance avec des valeurs minimales et maximales des différentes mesures, remplaçant — pour cette évaluation — les intervalles de confiance calculés au moyen de la simulation de Monte Carlo.

7.3.1 Évaluation multi-catégories

Nous présentons l'évaluation globale de Wapiti sur le corpus de test dans le tableau 7.5 et l'évaluation détaillée sur chaque catégorie dans le tableau 7.6.

Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure	Intervalles de confiance
572	38	82	0,8746	0,9377	0,9051	[0,8882;0,9220]

TABLE 7.5 – Évaluation globale de Wapiti sur le corpus de test en cardiologie

Catégorie	Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure
Dates	229	2	9	0,962	0,991	0,977
Noms	183	18	22	0,893	0,910	0,901
Prénoms	98	14	11	0,899	0,875	0,887
Hôpitaux	33	2	10	0,767	0,943	0,846
Villes	12	1	10	0,545	0,923	0,686
Codes postaux	5	0	3	0,625	1,000	0,769
Adresses	1	0	7	0,125	1,000	0,222
Téléphones	6	1	2	0,750	0,857	0,800
Appareillage	4	0	6	0,400	1,000	0,571
Numéro de série	1	0	2	0,333	1,000	0,500

TABLE 7.6 – Évaluation détaillée de Wapiti sur le corpus de test en cardiologie

7.3.2 Évaluation avec fusion nom/prénom

Nous présentons dans les tableaux 7.7 et 7.8 les résultats obtenus par Wapiti sur le corpus de test de cardiologie après avoir rassemblé les catégories *nom* et *prénom* sous une seule catégorie.

Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure	Intervalles de confiance
490	18	66	0,8813	0,9646	0,9211	[0,9043;0,9379]

TABLE 7.7 – Évaluation globale de Wapiti sur le corpus de test en cardiologie après avoir rassemblé les catégories nom et prénom

Catégorie	Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure
Noms	199	13	17	0,921	0,939	0,930
Hôpitaux	33	0	10	0,767	1,000	0,868
Villes	13	2	9	0,591	0,867	0,703
Codes postaux	4	0	4	0,500	1,000	0,667

TABLE 7.8 – Évaluation détaillée de Wapiti sur le corpus de test en cardiologie après avoir rassemblé les catégories nom et prénom

Dans le tableau 7.8, nous ne renseignons que les lignes de résultats qui diffèrent de celles du tableau 7.6, ce qui permet de mettre en évidence les catégories pour lesquelles la fusion des catégories *nom* et *prénom* a eu des effets de bord, en l'occurrence les noms d'hôpitaux, les noms de villes et les codes postaux.

7.3.3 Évaluation par validation croisée

Modèle global

Nous renseignons dans le tableau 7.9 les F-mesures calculées sur les résultats produits par Wapiti globalement et sur chacune des catégories au moyen de la validation croisée en 10 parties. À titre d'intervalle de confiance, nous avons relevé les valeurs minimale et maximale des F-mesures calculées parmi les dix parties de la validation croisée que nous indiquons comme valeurs de variation minimale et maximale de la F-mesure dans le tableau 7.9.

	Utilisation isolée des types			Combinaison de types		
	Caract. de surface	*Ressources externes*	*Formes des tokens*	*Tokens + Surface*	*Tokens + Ressources*	*Tokens + Surface + Ressources*
Rappel global	0,0200	0,8021	0,8280	0,8539	0,8662	**0,9012**
Précision globale	1,0000	0,8805	0,9467	**0,9537**	0,9519	0,9469
F-mesure globale	0,0500	0,8395	0,8834	0,9010	0,9070	**0,9235**
Variation min/max	0,0000	0,8237	0,8561	0,8836	0,8926	0,9089
de la F-mesure	0,4400	0,8681	0,9025	0,9212	0,9275	0,9383
Dates	0,0000	0,9270	0,9368	0,9484	0,9526	**0,9591**
Noms	0,1100	0,8104	0,8674	0,8910	0,8963	**0,9208**
Hôpitaux	0,0000	0,7081	0,8795	**0,8914**	0,8860	0,8910
Villes	0,0000	0,5610	0,6619	0,7040	0,7677	**0,8135**
Codes postaux	0,0000	0,7547	0,5128	0,5128	0,6512	**0,8163**
Adresses	0,0000	0,2778	0,2143	**0,4865**	0,3333	0,4138
Téléphones	0,0000	0,9649	**0,9823**	**0,9823**	**0,9823**	0,9778
Appareillages	0,0000	**0,2553**	0,1455	0,1455	0,1481	0,2456
Numéro de série	0,0000	0,0769	**0,9302**	0,9048	0,9091	0,8500

TABLE 7.9 – F-mesures obtenues par Wapiti en validation croisée en 10 parties

Dans ces expériences, les catégories *nom* et *prénom* ont fait l'objet d'un rassemblement sous une seule catégorie. Nous avons lancé plusieurs expériences d'anonymisation, en distinguant le type de caractéristique utilisé, soit en se restreignant à un seul type de caractéristiques *(colonnes centrales du tableau)*, soit en combinant plusieurs types de caractéristiques *(dernières colonnes du tableau)*. Le détail des F-mesures par catégorie pour chacune des six expériences [1] menées par validation croisée est également donné dans le graphique 7.1.

1. Les six expériences sont légendées comme suit : (i) Surface=caractéristiques de surface uniquement ; (ii) Externes=ressources externes uniquement ; (iii) Token=forme des tokens uniquement ; (iv) Tok+Surf=combinaison forme des tokens et caractéristiques de surface ; (v) Tok+Ext=combinaison forme des tokens et ressources externes ; (vi) Tok+Ext+Surf=combinaison maximale.

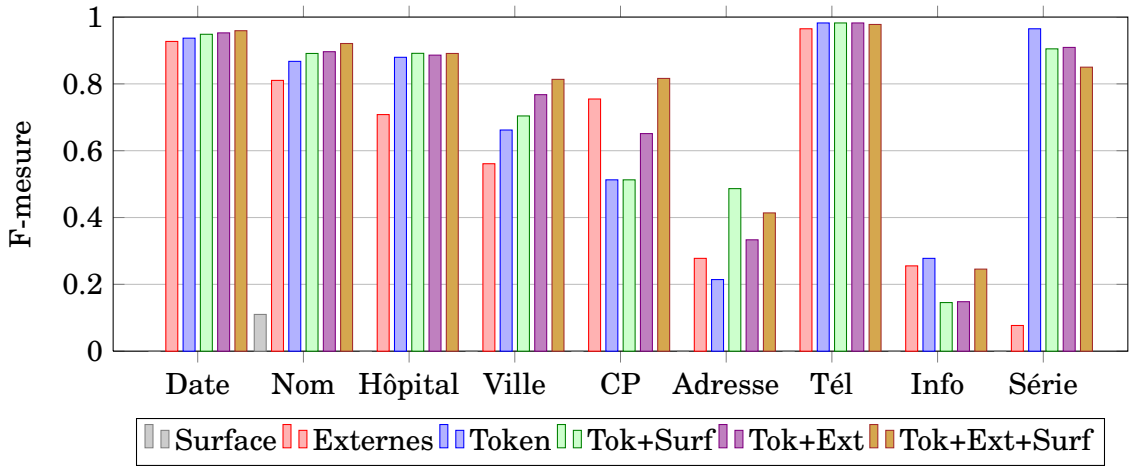


FIGURE 7.1 – Évolution des résultats par catégorie avec la validation croisée en 10 parties

Nous renseignons dans les tableaux 7.10 et 7.11 les résultats obtenus sur le corpus de test par l'application du modèle créé par Wapiti en validation croisée en 10 parties reposant sur la combinaison de tous les types d'information *(dernière expérience du tableau 7.9)*. Nous observons sur ce corpus des résultats similaires à ceux obtenus en validation croisée.

Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure	Intervalles de confiance
604	23	81	0,8818	0,9633	0,9207	[0,9056;0,9359]

TABLE 7.10 – Évaluation globale de Wapiti (modèle créé en validation croisée en 10 parties) sur le corpus de test en cardiologie après avoir rassemblé les catégories nom et prénom

Catégorie	Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure
Dates	222	2	16	0,933	0,991	0,961
Noms	324	15	21	0,939	0,956	0,947
Hôpitaux	30	1	13	0,698	0,968	0,811
Villes	11	1	11	0,500	0,917	0,647
Codes postaux	4	0	4	0,500	1,000	0,667
Adresses	1	2	7	0,125	0,333	0,182
Téléphones	6	1	2	0,750	0,857	0,800
Appareillage	5	0	5	0,500	1,000	0,667
Numéro de série	1	1	2	0,333	0,500	0,400

TABLE 7.11 – Évaluation détaillée de Wapiti (modèle créé en validation croisée en 10 parties) sur le corpus de test en cardiologie après avoir rassemblé les catégories nom et prénom

Modèles spécifiques par catégorie

Nous renseignons dans le tableau 7.12 les résultats obtenus par Wapiti, sur la base d'un modèle spécifique à chacune des catégories traitées, en utilisant le fichier de configuration dite « optimale » *(fichier utilisé pour la construction du modèle global)*. Les modèles ont tous été construits indépendamment les uns des autres, sur la base d'un corpus annoté uniquement avec la catégorie du modèle concerné.

Catégorie	Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure
Dates	1 511	50	137	0,9169	0,9680	0,9417
Noms	1 645	109	250	0,8681	0,9379	0,9016
Hôpitaux	268	9	65	0,8048	0,9675	0,8787
Villes	85	18	93	0,4775	0,8252	0,6050
Codes postaux	7	0	22	0,2414	1,000	0,3889
Adresses	1	0	22	0,0435	1,000	0,0833
Téléphones	109	1	5	0,9561	0,9909	0,9732
Appareillage	3	1	45	0,0625	0,7500	0,1154
Numéro de série	0	0	23	0,000	0,000	0,000

TABLE 7.12 – Évaluation détaillée de Wapiti en validation croisée avec un modèle par catégorie, configuration dite « optimale » d'utilisation des caractéristiques

Catégorie	Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure
Noms	1 654	107	241	**0,8728**	**0,9392**	**0,9048**
Hôpitaux	270	12	63	**0,8108**	0,9574	0,8780

TABLE 7.13 – Évaluation détaillée de Wapiti en validation croisée avec un modèle par catégorie, configuration améliorée d'utilisation des caractéristiques

Nous indiquons dans le tableau 7.13 l'évolution des résultats après avoir testé l'amélioration du fichier de configuration d'utilisation des caractéristiques produites. Parmi nos différentes tentatives d'amélioration, seules deux catégories présentent des résultats intéressants : les noms de personne et les noms d'hôpitaux.

Sur les noms de personne, nous avons ajouté un patron qui tient compte de la présence du token dans un syntagme nominal. L'utilisation de cette règle nous permet de traiter correctement neuf noms supplémentaires et de dépasser les résultats obtenus avec la configuration dite « optimale ».

Sur les noms d'hôpitaux, nous avons modifié la configuration des caractéristiques issues de l'étiquetage en parties du discours : suppression des trigrammes d'étiquettes et ajout de l'étiquette du token qui suit. Cette configuration permet uniquement d'augmenter le rappel, avec deux noms corrects supplémentaires pour les noms d'hôpitaux. Mais elle présente l'inconvénient de sur-anonymiser et réduit en conséquence la précision et la F-mesure. Un processus de sur-anonymisation qui permet d'augmenter le rappel est néanmoins préférable dans une perspective d'anonymisation et de confidentialité des données.

7.4 Enchaînement de méthodes

Enfin, nous donnons dans le tableau 7.15 les résultats obtenu par l'enchaînement des deux outils utilisés dans nos différentes expériences : d'abord Wapiti en validation croisée en 10 parties pour les approches à base d'apprentissage *(dont les résultats sont donnés dans les tableaux 7.9 et 7.14)*, suivi de Medina pour les méthodes symboliques. Nous précisons que Medina a été appliqué sur les sorties produites par Wapiti.

Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure	Intervalles de confiance
4 033	422	263	0,9388	0,9053	0,9217	[0,9159;0,9276]

TABLE 7.14 – Évaluation globale de Wapiti en validation croisée suivi de Medina

Catégorie	Vrais Positifs	Faux Positifs	Faux Négatifs	Rappel	Précision	F-mesure
Dates	1 595	71	53	0,9678	0,9574	**0,9626**
Noms	1 809	180	90	0,9526	0,9095	**0,9306**
Hôpitaux	290	121	44	0,8683	0,7056	0,7785
Villes	146	28	32	0,8202	0,8391	**0,8295**
Codes postaux	26	5	3	0,8966	0,8387	**0,8667**
Adresses	16	1	7	0,6957	0,9412	**0,8000**
Téléphones	110	2	4	0,9649	0,9821	0,9735
Appareillage	18	13	30	0,3750	0,5806	**0,4557**
Numéro de série	23	0	0	1,000	1,000	**1,000**

TABLE 7.15 – Évaluation détaillée de Wapiti en validation croisée suivi de Medina

7.5 Analyse des erreurs

7.5.1 Sur la robustesse du système symbolique

Le système Medina fondé sur les méthodes symboliques a été développé en prenant pour exemple les données sur lesquelles il se destinait à être appliqué : des comptes rendus de cardiologie du CHU Pontchaillou à Rennes. En conséquence, les règles et les ressources ont été produites en se fondant sur les caractéristiques des cas rencontrés en corpus. L'utilisation de Medina sur des comptes rendus cliniques du CHU de Lille n'a pas permis d'anonymiser les adresses de messagerie électronique parce que ce type d'information n'était pas présent dans les données de Rennes, et qu'aucune règle n'avait donc été envisagée pour les traiter avant de les rencontrer.

Mais l'application de l'outil Medina sur un nouveau corpus de données risque également de générer du bruit, parce que les règles n'auront pas été conçues ni adaptées aux caractéristiques de ces nouvelles données. Nous avons ainsi pu observer que l'application de Medina sur un corpus de comptes rendus en oncologie de l'Hôpital Européen George Pompidou a produit quelques anonymisations erronées. La règle qui identifie un nom d'hôpital fondée sur la présence du déclencheur « centre » [2] a été

2. A côté des déclencheurs *hôpital, clinique, CHU (indices « pertinents »)*, certains indices « ambigus »

déclenchée de manière erronée sur ce corpus (exemple 20). Faute d'avoir rencontré ce type de construction sur le corpus de Rennes, il ne nous a pas été possible d'identifier et de corriger cette erreur avant d'appliquer Medina sur un nouveau jeu de données.

(20) *Présence d'adénopathies médiastinales centimétriques avec* [hôpital] *centre graisseux d'allure banale*.

7.5.2 Sur l'enchaînement des méthodes

Nous avons procédé à une analyse des erreurs produites lors de la dernière expérience. Cette expérience repose sur l'enchaînement des deux méthodes *(apprentissage statistique puis méthodes symboliques)* avec une approche statistique reposant sur la construction et l'application d'un modèle par validation croisée en 10 parties. Nous avons retenu cette expérience car il s'agit de l'expérience qui nous a permis d'obtenir les meilleurs résultats sur la majorité des catégories (tableau 7.15) et notamment sur la catégorie des noms de personnes qui nous intéresse plus particulièrement, du fait de son importance en matière d'anonymisation et de respect de la vie privée.

Sur la catégorie des noms de personnes, nous notons que 52 occurrences ont été manquées par l'enchaînement des deux outils. Parmi ces 52 occurrences, nous relevons trois types de faux négatifs :

- Un faux négatif correspond à l'initiale isolée d'un prénom : *P.* ;
- Quinze faux négatifs (29 %) sont dus à des abréviations de noms et prénoms présentes de manière isolée sur une ligne, sans qu'il ne soit possible de s'appuyer sur le contexte immédiat pour définir des règles : *ALH / MHF, DB, HLB* ;
- La majorité des faux négatifs (36, soit 69 %) correspond à l'un des cas suivants :
  - Des noms et prénoms réels absents des listes utilisées : *Adonis, Nenci* ;
  - Des erreurs typographiques : soit l'oubli d'une espace entre un nom et un prénom, ou entre un titre et un nom (*PrMARTIN*), soit l'utilisation du chiffre zéro « 0 » au lieu de la voyelle « O » en majuscule (*0. Martin*), conduisant les systèmes à ne pas déclencher les règles implémentées.

7.6 Discussion

7.6.1 Comparaison des approches

Méthodes symboliques : des résultats proches de l'état de l'art

Conformément aux observations réalisées sur les résultats d'expériences d'anonymisation à base de méthodes symboliques (voir section 2.2, et la section 1.5.2 pour les résultats obtenus sur le défi i2b2 2006), notre système Medina obtient une précision supérieure au rappel. Ces résultats confirment qu'en utilisant des méthodes symboliques, il est possible d'obtenir des résultats de qualité *(bonne précision, 0,8630)* mais avec des difficultés de couverture *(rappel plus faible, 0,8328)*.

Nous observons par ailleurs que ces résultats s'inscrivent dans la lignée de ceux que nous avions obtenus en 2002, en travaillant à l'anonymisation d'un corpus de stomatologie, pour lequel la précision se révélait également supérieure au rappel. Les résultats se révèlent même particulièrement proches *(Stomato, F=0,8478 ; Medina, F=0,8476 sur l'ensemble des catégories et F=0,8606 après fusion nom/prénom)*, malgré les différences de tailles et de caractéristiques des corpus utilisés (section 5.2.3).

tels que *centre* ont été envisagés sans qu'ils n'aient jamais été utilisés sur les données de Rennes.

Ayant utilisé deux systèmes différents, tous deux adaptés au domaine pour lequel ils ont été conçus, nous ne pouvons tirer de conclusion quant au domaine médical *(et ses spécificités linguistiques et structurelles sous-jacentes)* sur lequel est appliqué un système et sur les conséquences qui résultent d'une telle application. Nous constatons toutefois que l'intervalle de confiance calculé sur la F-mesure globale de Medina est assez réduit [0,8266;0,8687], ce qui suggère que la qualité d'anonymisation serait maintenue en cas de passage à l'échelle. Nous pouvons donc en conclure que l'anonymisation réalisée sur le corpus complet de 21 749 documents est d'une qualité similaire à celle obtenue sur le corpus de test composé de seulement 62 documents.

Il reste toutefois des marges de progression, notamment au regard des performances atteintes par MedTag [Ruch et al., 2000] sur le français avec des taux d'anonymisation avoisinant les 98 / 99 % dans les catégories traitées. Nous estimons qu'introduire des informations morpho-syntaxiques de meilleure qualité pourrait permettre d'améliorer les résultats actuels de Medina. En effet, nous considérons que la définition de patrons fondés sur les parties du discours des tokens permettrait de traiter plus efficacement les longues entités telles que les adresses postales ou les noms d'hôpitaux, deux catégories que l'approche à base de règles n'a pas réussi à traiter efficacement. En travaillant au niveau des parties du discours et non plus directement sur la reconnaissance de tokens, nous pensons que l'outil devrait être en mesure d'identifier davantage d'informations relevant de ces deux catégories.

Méthodes par apprentissage : avantage net

Dès lors que la base annotée fournie à l'apprentissage est de qualité *(tant du point de vue des annotations réalisées par les humains que de celui de la représentativité des informations à traiter)*, les méthodes par apprentissage statistique permettent d'obtenir globalement de bons résultats.

Si l'utilisation seule de la forme des tokens permet d'obtenir une F-mesure de 0,8834 (tableau 7.9), déjà supérieure à celle obtenue par les méthodes symboliques, l'utilisation combinée avec les caractéristiques de surface autorise un gain de deux points, portant la F-mesure à 0,9010.

Méthodes hybrides : un gain réel

Le choix des types de caractéristiques à utiliser lors de la construction du modèle constitue donc un autre point essentiel dans la réussite des méthodes statistiques. Si l'utilisation de la forme des tokens combinée avec des caractéristiques de surface permet d'atteindre une F-mesure de 0,9010 (tableau 7.9), la combinaison de ces informations avec des propriétés linguistiques issues de ressources externes et produites par des outils du traitement automatique des langues *(étiquetage en parties du discours, présence du token en lexique ou parmi une liste de déclencheurs, regroupement automatique des tokens)* permet encore de gagner deux points supplémentaires de F-mesure à 0,9235.

Sur le corpus de test (figure 7.2), Wapiti obtient ainsi de bien meilleurs résultats que Medina, qu'il s'agisse des anonymisations réalisées avec conservation des catégories *nom* et *prénom* (Medina MC : F=0,8476 ; Wapiti MC : F=0,9051) ou après avoir regroupé ces catégories en une seule catégorie globale (Medina F : F=0,8606 ; Wapiti F : F=0,9211). On observe le même écart de 6 points de F-mesure environ entre les deux outils. La validation croisée en 10 parties confirme ces résultats (Wapiti VC : F=0,9235).

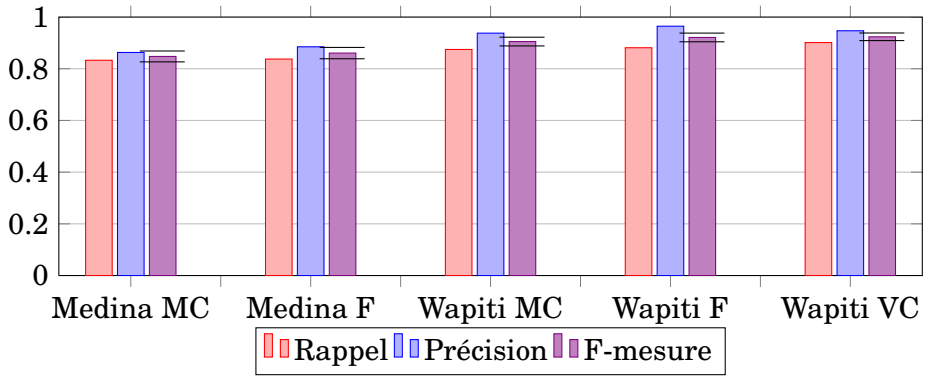


FIGURE 7.2 – Évaluations globales de Medina et Wapiti ; sur le corpus de test en multi-catégories (MC) ou après fusion nom/prenom (F) ; en validation croisée (VC). Intervalles de confiance renseignés pour la F-mesure

Les intervalles de confiance calculés sur la F-mesure, symbolisés par des traits noirs sur la figure 7.2, se révèlent assez réduits *(+/-2,2 points de F-mesure pour Medina et +/-1,7 points pour Wapiti)*. Ces intervalles démontrent que l'utilisation des outils sur un plus grand nombre de fichiers produirait des résultats de qualité similaire.

Un modèle global plus efficace que la somme de modèles spécifiques

Nous constatons que l'application d'un modèle global, c.-à-d. un modèle unique permettant de traiter l'ensemble des catégories d'information à anonymiser, obtient de bien meilleurs résultats (tableau 7.9) que l'utilisation de modèles spécifiques à chacune des catégories (tableau 7.12). En effet, aucun modèle propre à une catégorie donnée ne surpasse le modèle globalement créé. Cette observation confirme que les CRF de chaînes linéaires sont efficaces avec un modèle global plutôt qu'un ensemble de modèles spécifiques qui risquent de prendre des décisions incompatibles.

Si la configuration optimale du modèle global se révèle également optimale pour la majorité des catégories, certaines catégories bénéficient de l'ajout ou de la suppression de patrons dans le fichier de configuration utilisé lors de la construction du modèle. Ainsi, sur la catégorie des noms, la configuration optimale obtient une F-mesure de 0,9016 alors qu'en ajoutant la règle qui tient compte de la présence du token dans un syntagme nominal, elle monte à 0,9048 *(9 noms corrects supplémentaires ont ainsi été trouvés par l'ajout de cette règle)*, sans pour autant dépasser la F-mesure obtenue avec le modèle global *(F=0,9208)*. Sur les autres catégories, la configuration optimale du modèle global reste la meilleure.

Sur ces expériences de réalisation de modèles spécifiques, nous pouvons donc en conclure que le traitement d'une catégorie donnée nécessite, en plus des caractéristiques produites *(formes des tokens, caractéristiques de surface, et ressources externes)*, des informations relatives aux autres catégories. La construction d'un modèle prend donc en compte la totalité des informations du corpus, ce qui lui assure une meilleure robustesse.

Des résultats inférieurs à l'anonymisation humaine

Nous observons par ailleurs que les résultats obtenus aussi bien au moyen des méthodes symboliques qu'avec les méthodes par apprentissage statistique se révèlent inférieurs à ceux obtenus grâce à une annotation humaine.

En section 4.3.3, nous avons calculé la F-mesure obtenue par chacun des deux annotateurs humains vis à vis du résultat de la fusion de cette double annotation. Nous rappelons que l'annotation et l'évaluation humaine ont porté sur le corpus avec les catégories *nom* et *prénom*. Nous donnons ci-après les résultats obtenus par les deux approches avec ces deux catégories traitées.

Le premier annotateur obtenait une F-mesure de 0,8698 contre 0,9307 pour le second. Bien que l'évaluation sur le corpus de test ne porte que sur 62 des 100 fichiers annotés en double, Medina obtient une F-mesure moindre, s'élevant à 0,8476.

En matière d'apprentissage statistique, Wapiti obtient une meilleure F-mesure que le premier annotateur, mais il ne parvient pas à dépasser le second, avec une F-mesure finale de 0,9051.

7.6.2 Une qualité inégale selon les catégories

Dans le détail, nous observons que certaines catégories ont fait l'objet d'un meilleur traitement que d'autres (figure 7.3). Ces performances doivent toutefois être appréciées à l'aune du nombre d'entités présentes dans chaque catégorie (voir tableau 4.1).

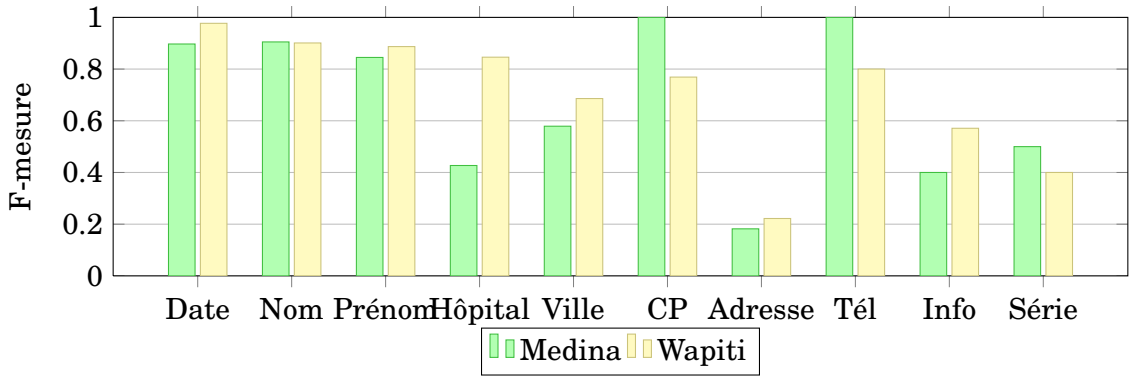


FIGURE 7.3 – Évaluation de Medina et Wapiti sur chaque catégorie avant fusion

Méthodes symboliques : bonne qualité des données numériques et nominatives simples

Si les données numériques *(codes postaux, dates, téléphones, numéro de série)* permettent d'obtenir assez aisément de bons résultats avec les méthodes symboliques *(en raison des possibilités plus simples de représentation du format des informations)*, il en est tout autre des données nominatives. Les données reposant sur la projection de lexiques, éventuellement complétée de quelques règles, permettent d'obtenir de bons résultats *(noms, prénoms)*. En revanche, le traitement des informations nominatives relevant essentiellement de règles *(adresses postales, villes, noms d'hôpitaux)* se révèle particulièrement complexe, en raison de la difficulté à définir des règles efficaces et génériques.

Méthodes par apprentissage : bonne qualité des données bien représentées

Conformément à nos prévisions, nous constatons que les méthodes par apprentissage ont moins bien réussi que les méthodes symboliques sur les données numériques *(codes postaux, téléphones, numéro de série)*, à l'exception de la catégorie des dates dont le nombre élevé d'entités a permis à l'outil CRF de construire un modèle robuste sur cette catégorie d'information.

Puisque les approches à base d'apprentissage statistique ne cherchent pas à construire des règles mais à construire des modèles fondés sur des caractéristiques surfaciques et contextuelles partagées par les entités d'une catégorie, les données complexes qui n'ont pu être traitées correctement par les méthodes symboliques *(noms d'hôpitaux ou villes)* l'ont été par l'apprentissage. L'approche à base de règles aura, soit déclenché une règle pour une portion ne nécessitant pas de traitement (exemple 21), soit attribué une mauvaise étiquette sur une portion correcte (exemple 22). L'approche par apprentissage statistique aura correctement identifié les frontières et la catégorie sur ce même exemple (exemple 23). La catégorie des adresses postales, faiblement représentée dans le corpus, constitue un point fortement négatif, quelle que soit l'approche envisagée. Ainsi, les adresses ne sont généralement pas anonymisées.

(21) *L'examen* [hôpital *clinique cardio-vasculaire demeure*] *normal*

(22) *unité* [prénom *P*] [nom *Ledoyen*]

(23) [hôpital *unité P Ledoyen*]

Nous estimons toutefois que des expériences permettant de biaiser le rappel devraient permettre d'identifier davantage de données. Tout outil d'apprentissage statistique permet d'associer un score à chaque prédiction. La modification du seuil défini par défaut en abaissant la valeur de ce seuil permettrait de prendre en compte davantage de prédictions, donc de biaiser le rappel. Ces expériences restent à être menées.

7.6.3 Une fusion nom/prénom bénéfique

Nous observons que le regroupement des catégories *nom* et *prénom* en une seule catégorie permet d'améliorer globalement les résultats (figure 7.4), tant pour les approches symboliques que pour les approches à base d'apprentissage. Les deux approches obtiennent toutes deux une F-mesure de 0,93 sur cette nouvelle catégorie.

Méthodes symboliques : une augmentation mécanique

Medina ayant été conçu pour traiter deux catégories *nom* et *prénom* distinctes, l'outil se sert des informations d'appartenance à l'une ou l'autre de ces deux catégories pour effectuer des anonymisations complémentaires[3] *(les patrons syntaxiques définis dans le programme reposent sur la présence spécifique de ces deux catégories nom et prénom).* En conséquence, nous n'avons rien modifié de l'outil pour qu'il ne traite qu'une seule catégorie. Le regroupement des catégories *nom* et *prénom* a donc

3. Une lettre isolée en majuscule suivie d'un élément étiqueté *nom* sera étiquetée *prénom* : par exemple, *A.* <*nom*>.

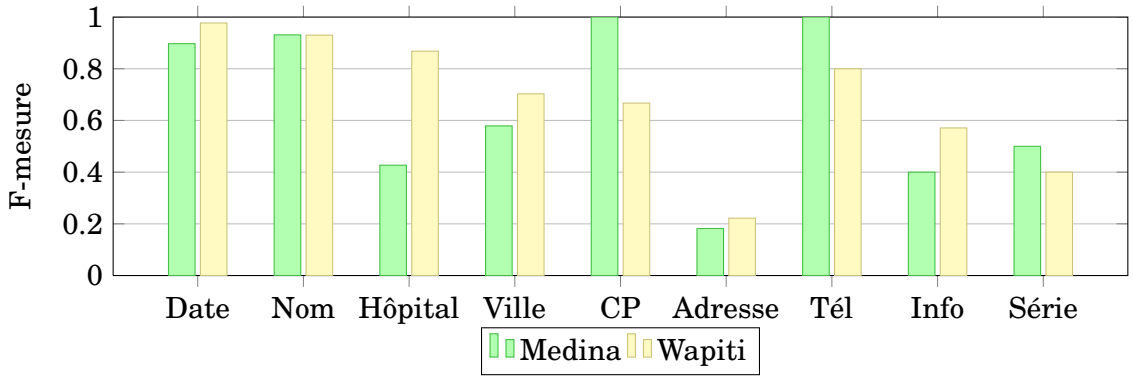


FIGURE 7.4 – Évaluation de Medina et Wapiti après fusion nom/prénom

été réalisé lors de l'évaluation. Les résultats obtenus par regroupement après coup ne présagent donc rien des capacités du systèmes à mieux traiter une seule catégorie globale issue du regroupement de deux catégories que deux catégories séparées.

Nous observons tout d'abord que la F-mesure globale augmente, passant de 0,8476 à 0,8606 avec un gain de 1,3 point. Dans le détail, nous constatons également que les F-mesures obtenues sur ces deux catégories *(noms=0,905 et prénoms=0,845)* sont inférieures à celle obtenue sur la catégorie unique qui les regroupe *(noms=0,931)*. Toutefois, ces augmentations sont uniquement mécaniques, d'une part sous l'effet de la fusion *(les erreurs de typage entre nom et prénom disparaissent)*, d'autre part au bénéfice du nombre plus réduit d'entités à trouver dans cette nouvelle catégorie *(205 noms et 109 prénoms qui donnent lieu à 216 noms après fusion)*.

Méthodes par apprentissage : une amélioration avec effets de bord positifs

Nous constatons que cette fusion de catégories a également été bénéfique pour les méthodes par apprentissage. Le premier bénéfice concerne l'augmentation de la F-mesure globale, qui passe de 0,9051 à 0,9211 avec un gain de 1,6 points, légèrement supérieur à celui obtenu par Medina. Comme pour les méthodes symboliques, la fusion conduit à l'obtention d'une F-mesure plus élevée sur cette nouvelle catégorie *(noms=0,930)* que sur les deux catégories d'origine *(noms=0,901 et prénoms=0,887)*.

Le second bénéfice engendré par cette fusion concerne des effets de bord sur trois autres catégories. Ces effets sont positifs sur deux catégories *(hôpitaux et villes)* et conduisent à une augmentation de la F-mesure. En revanche, les résultats d'une troisième catégorie *(codes postaux)* s'en trouvent dégradés. Nous pensons que ces modifications de résultats s'expliquent par le fait que les noms d'hôpitaux *(et plus encore les services et unités à l'intérieur d'un hôpital)* sont généralement constitués d'un prénom et d'un nom, en hommage à une personne. Sur la catégorie des noms de villes, nous estimons que la proximité dans le contexte de ces occurrences d'une adresse postale avec un nom de voirie également composé du prénom et du nom d'une personne permet d'expliquer cette amélioration. La baisse de qualité sur la catégorie des codes postaux est peut-être due, lors de la construction du modèle, à la création d'un nombre plus réduit de règles du fait de la réorganisation des différentes entités. Il est cependant difficile de ne pas se perdre en conjectures pour tenter de trouver une explication à ces changements.

7.6.4 Un enchaînement de méthodes positif

Enfin, l'enchaînement des approches par apprentissage suivies des méthodes symboliques a permis au système à base de règles *(Medina)* de compléter les traitements effectués par le système par apprentissage statistique *(Wapiti)*. Nous observons que cette « cascade de systèmes » a permis l'augmentation des résultats obtenus sur sept catégories d'informations parmi un total de neuf. Si la F-mesure globale n'augmente pas, on observe en revanche une inversion des valeurs de rappel et de précision. Autrement dit, l'enchaînement des deux approches a permis d'identifier davantage d'informations à anonymiser, conduisant à un rappel plus élevé *(R=0,9388 en enchaînement contre R=0,9012 par l'apprentissage seul)*, que chaque méthode prise séparément, ce qui est particulièrement appréciable pour les objectifs de l'anonymisation.

7.7 Bilan sur les informations « sensibles »

Nous terminons ce chapitre par un bilan du traitement qui a été appliqué aux informations les plus sensibles du corpus, en l'occurrence les informations nominatives sur les patients *(noms et prénoms)*. Si, en matière d'anonymisation, une certaine tolérance peut s'appliquer à certaines catégories d'informations qui n'auraient pas été traitées (section 1.4.1), les informations permettant d'identifier un patient doivent impérativement être rendues anonymes. L'approche que nous avons suivie dans ce travail de thèse repose sur deux étapes.

7.7.1 Appariement avec le Système d'Information Patient.

La première étape d'anonymisation a consisté à effectuer un appariement à l'identique entre le contenu du Système d'Information Patient (SIP) de l'hôpital et les fichiers du corpus (section 4.3.2).

Nous avons étudié les 562 fichiers de nos trois corpus (section 4.3.3) dans la version qui nous a été fournie par le CHU partenaire, c.-à-d. ayant fait l'objet de cette première étape d'anonymisation. Nous observons ainsi qu'à l'issue de cette première étape, 25 % des fichiers contiennent toujours le nom ou le prénom du patient, ou d'un membre de la famille du patient (voir tableau 7.16). Nous relevons également que l'absence de traitement de ces informations nominatives concerne davantage les prénoms (75 % des cas) que les noms (25 % des cas).

Corpus	Appr.	Dev.	Test
Nombre total de fichiers	250	250	62
Fichiers avec nom/prénom du patient	58 (23,2 %)	62 (24,8 %)	19 (30,6 %)
Nombre total de lignes	8 030	9 497	2 087
Lignes avec nom/prénom du patient	85 (1,06 %)	87 (0,92 %)	21 (1,01 %)
Nombre de noms non anonymisés	22	25	6
Nombre de prénoms non anonymisés	71	72	20

TABLE 7.16 – Statistiques sur les informations patients non anonymisées au terme de la première étape (appariement avec le SIP de l'hôpital) dans les trois corpus utilisés

Cette absence de traitement correspond à l'un des cas suivants :
- les noms et prénoms comportant une lettre avec un accent *(Céline, Marie-Hélène)* ou un autre diacritique *(François)* ;
- les noms et prénoms composés *(Anne-Marie, Jean-Marc)* et les prénoms collés à une apostrophe : *« l'état cardiaque d'Anna »* ;
- les noms et prénoms intégralement écrits en majuscules : *« Monsieur MARTIN se plaint toujours des mêmes douleurs »* ;
- les noms et prénoms des membres de la famille du patient, forcément absents du SIP : *« une surdité acquise chez sa sœur Marie »*, *« les ECG de surface des deux enfants (Audrey et Benjamin) ne montrent pas d'anomalie »* ;
- les autres cas pour lesquels, faute de pouvoir accéder au SIP de l'hôpital, nous n'avons pas d'explication sur l'absence de réalisation d'anonymisation *(ni accent, ni prénom composé, ni réalisation en capitales)* : *Caroline, Daniel, Karim*.

Il apparaît globalement que l'appariement à l'identique a été réalisé de manière beaucoup trop stricte, et n'autorise aucun écart *(encodage, casse typographique)* par rapport à la version renseignée dans le SIP.

7.7.2 Anonymisation complémentaire.

La deuxième étape repose sur l'utilisation complémentaire d'un système d'aide à l'anonymisation automatique reposant, soit sur des méthodes symboliques au moyen de l'outil Medina (section 5.4), soit sur des approches par apprentissage statistique via les CRF de chaînes linéaires (section 6.3), soit par l'hybridation des deux types de méthode précédents notamment par l'enchaînement de ces deux méthodes (section 7.4). Nous avons pu constater que l'enchaînement des méthodes a permis un meilleur traitement général des données que chaque méthode prise isolément, et tout particulièrement sur la catégorie des noms *(F-mesure=0,9306).* [4]

Nous avons également analysé le résultat de l'anonymisation effectuée par ces différents outils, du point de vue des informations nominatives relatives au patient ou aux membres de la famille du patient.

Méthodes par apprentissage statistique

Nous avons évalué la capacité du système par apprentissage statistique Wapiti a traiter les noms et prénoms des patients laissés en clair par la première étape d'anonymisation. Cette évaluation est fondée sur les sorties générées par le système, sur la base d'un modèle global traitant toutes les catégories et notamment en regroupant les catégories *nom* et *prénom* en une seule catégorie. L'expérience que nous avons retenue étant fondée sur une validation croisée, le corpus traité intègre les fichiers des corpus d'apprentissage et de développement, soit un total de 500 fichiers.

Nous renseignons dans le tableau 7.17 le nombre d'informations nominatives relatives au patient qui demeurent en clair à l'issue de la double étape d'anonymisation *(d'abord l'anonymisation par appariement avec le SIP puis l'application du système par apprentissage).*

La comparaison des tableaux 7.17 et 7.16 nous permet de mettre en évidence une forte baisse du nombre d'informations nominatives relatives au patient qui demeurent en clair à l'issue du double processus d'anonymisation. Cependant, il reste

4. Pour rappel, la catégorie « noms » sur nos dernières expériences regroupe, d'une part les prénoms et les noms, d'autre part, les informations nominatives sur le patient mais également sur les membres de l'équipe médicale. La F-mesure que nous indiquons ici se rapporte donc à cette catégorie générique.

Corpus	Appr.	Dev.	Test
Nombre total de fichiers	250	250	62
Fichiers avec nom/prénom du patient	4 (1,6 %)	6 (2,4 %)	4 (6,5 %)
Nombre total de lignes	8 030	9 497	1 911
Lignes avec nom/prénom du patient	4 (0,05 %)	9 (0,09 %)	6 (0,31 %)
Nombre de noms non anonymisés	0	2	4
Nombre de prénoms non anonymisés	4	12	0

TABLE 7.17 – Statistiques sur les informations patients non anonymisées au terme de la deuxième étape (Wapiti) dans les deux corpus utilisés en validation croisée et sur le corpus de test (application du modèle créé en validation croisée)

un petit nombre d'informations qui ne permettent pas de garantir la confidentialité des données par ce seul traitement statistique.

Sur les 500 fichiers des deux corpus traités, 10 fichiers (2 %) sont encore porteurs d'informations personnelles, parmi lesquelles figurent deux occurrences du même nom propre d'un patient *(intégralement écrit en majuscules)*. En ce qui concerne les prénoms non traités, nous observons qu'ils figurent fréquemment en tout début de ligne *(4 occurrences sur 16)* ou entre parenthèses *(5 occurrences sur 16)*.

Le modèle CRF que nous avons construit semble donc peu robuste à ces contextes particuliers. L'application du modèle créé en validation croisé en 10 parties sur le corpus de test confirme cette absence de robustesse. Nous observons que tous les prénoms des patients ont été traités alors que quatre noms de famille *(un nom apparaissant trois fois)* ne sont pas anonymisés. Aucun de ces noms n'est cependant un nom réel, il s'agit de noms réintroduits lors de la préparation des données. Ces noms ont fait l'objet d'une anonymisation lors de l'appariement du document avec le SIP.

Méthodes symboliques

L'application de l'outil à base de règles « Medina-RB » (section 5.4) sur les 62 fichiers du corpus de test a permis de traiter les noms et prénoms des patients dans leur intégralité, incluant donc ceux qui étaient restés en clair à l'issue de la première étape d'anonymisation *(20 prénoms et 6 noms)*.

Les résultats obtenus par notre système en termes de rappel sur la catégorie des noms et prénoms (R=0,907 sur les noms, R=0,927 sur les prénoms et R=0,963 par la fusion des catégories *nom*/*prénom*, section 7.2) et l'absence d'obtention d'un rappel de 1,000 sur ces catégories s'expliquent par deux raisons :

- les noms et prénoms qui n'ont pas été anonymisés à l'issue de cette procédure en deux étape sont quasi exclusivement ceux des chirurgiens de l'hôpital ou des médecins consultés par le patient avant son séjour hospitalier, soit sous forme pleine, sous forme d'initiales (section 7.5) ;
- de manière marginale, il peut s'agir de noms et de prénoms attribués à des patients lors de la procédure de réintroduction de noms (section 4.3.3), autrement dit, il s'agit de noms et prénoms fictifs. Ainsi, deux prénoms réintroduits *(Adonis, Nenci)* en remplacement des balises d'anonymisation de premier niveau n'ont pas été identifiés par Medina-RB parce qu'ils étaient absents des listes utilisées par l'outil.

En conclusion, la procédure d'anonymisation en deux étapes permet bien d'assurer un traitement à 100 % des noms et prénoms relatifs aux patients.

Enchaînement apprentissage statistique puis symbolique

Enfin, sur l'enchaînement des deux méthodes précédentes, le fait de terminer le processus d'anonymisation par les méthodes symboliques — qui ont permis de traiter toutes les données personnelles nominatives — permet effectivement de ne laisser en clair aucun nom ou prénom d'un patient.

7.8 Synthèse

Dans ce chapitre, nous avons exposé les résultats des différentes expériences réalisées. Au fur et à mesure de l'avancée des travaux d'anonymisation — en particulier lors des expériences en apprentissage statistique —, nous avons considéré que la distinction des catégories *nom* et *prénom* n'était pas nécessairement utile. Nous avons donc réalisé deux évaluations, la première portant sur la totalité des catégories initialement définies *(que nous avons appelée « évaluation multi-catégories »)*, la seconde après avoir rassemblé les noms et prénoms sous une seule catégorie *(présentée comme « évaluation après fusion nom/prénom »)*.

Méthodes symboliques. Nos expériences sur les méthodes symboliques ont permis l'obtention de résultats similaires à ceux présentés dans la littérature du domaine, avec des F-mesures globales de 0,8476 en multi-catégories et de 0,8606 après la fusion des catégories *nom* et *prénom*. Le regroupement de catégories ayant été réalisé lors de l'évaluation, cette augmentation de F-mesure n'est donc que mécanique, l'outil Medina n'ayant pas fait l'objet d'une modification quelconque pour regrouper ces deux catégories. Dans le détail, la F-mesure obtenue sur la catégorie générique *(0,931)* dépasse celles des deux catégories avant fusion *(0,905 sur les noms et 0,845 sur les prénoms)*, à relativiser du fait du nombre réduit d'instances au terme de la fusion *(205 noms et 109 prénoms avant fusion, 216 noms après fusion)*.

Approches par apprentissage statistique. Les approches à base d'apprentissage statistique ont conduit à l'obtention d'une F-mesure globale plus élevée, de 0,9051 en multi-catégories et de 0,9211 après la fusion. Comme pour les méthodes symboliques, la F-mesure obtenue sur la catégorie générique *(0,930)* est supérieure à celles des deux précédentes catégories *(0,901 sur les noms et 0,887 sur les prénoms)*, avec les mêmes précautions d'usage en termes de comparaison relatives au nombre d'instances différentes avant et après fusion. Nous avons également observé que le regroupement des deux catégories a produit des effets de bord majoritairement positifs, avec une augmentation de la F-mesure sur les noms d'hôpitaux *(passant de 0,846 avant fusion à 0,868 après fusion)* et les noms de villes *(passant de 0,686 avant fusion à 0,703 après fusion)*. À l'inverse, la F-mesure sur les codes postaux a diminué *(passant de 0,769 avant fusion à 0,667 après fusion)*. Faute de comprendre précisément le fonctionnement du CRF, nous estimons que la construction du modèle après fusion a conduit à la création de caractéristiques et poids différents dont les résultats se traduisent par ces effets de bord.

Après avoir mis en place un protocole expérimental reposant sur une validation croisée en 10 parties, nous avons mené plusieurs expériences reposant sur l'utilisation de différents types de caractéristiques utilisés pour construire le modèle de l'outil CRF : la forme des tokens tels qu'ils se présentent dans le document, les

caractéristiques de surface inférées de ces tokens *(casse typographique, présence de chiffres ou de ponctuation, taille)*, et des propriétés issues de ressources externes *(présence du token dans un lexique ou une liste de déclencheurs, étiquetage morpho-syntaxique, clusters)*. L'utilisation des seules caractéristiques de surface ne produit aucun résultat *(F=0,0500)*, les ressources externes uniquement permettent l'obtention de résultats proches de ceux des méthodes symboliques *(F=0,8375)* et l'utilisation de la forme des tokens uniquement les dépasse *(F=0,8834)*. Nous avons vu que l'utilisation combinée de plusieurs types de caractéristiques permet d'augmenter la valeur de la F-mesure et de gagner quatre points jusqu'à 0,9211.

La production de modèles spécifiques à chacune des catégories d'information à traiter ne permet pas d'améliorer les résultats. D'autre part, nous avons pu constater que le fichier de configuration précisant les conditions d'utilisation des caractéristiques produites tel qu'il a été utilisé pour la production du modèle global, qualifié de configuration « optimale », donne également les meilleurs résultats. La seule possiblité d'amélioration que nous avons atteinte concerne la catégories des noms pour lesquels l'information de présence du token dans un token nominal permet de gagner quelques points. Cependant, cette amélioration ne nous permet pas de dépasser les résultats que nous avons obtenus avec un seul modèle qui traite globalement de toutes les catégories.

Hybridation des méthodes. Enfin, l'enchaînement des approches par apprentissage suivies des méthodes symboliques a permis au système à base de règles *(Medina)* de compléter les traitements effectués par le système par apprentissage statistique *(Wapiti)*. Cet enchaînement a permis d'identifier davantage d'informations à anonymiser, conduisant à un rappel plus élevé *(R=0,9388 en enchaînement contre R=0,9012 par l'apprentissage seul)*.

Informations « sensibles ». En matière de traitement des informations les plus sensibles, en l'occurrence les noms et prénoms des patients, nous avons pu constater que la première étape d'anonymisation par appariement des informations avec le SIP de l'hôpital laisse environ 25 % des documents porteurs d'informations personnelles. La seconde étape d'anonymisation au moyen des méthodes symboliques (Medina-RB) permet de compléter cette première étape en anonymisant toutes les informations personnelles laissées en clair. A l'inverse, la deuxième étape fondée sur les approches par apprentissage statistique laisse encore quelques rares informations en clair dans 2 % des documents. Ainsi, si l'anonymisation par apparentissage statistique permet d'obtenir de meilleurs résultats que par le biais des méthodes symboliques, il apparaît sur nos données que le risque de laisser des informations personnelles en clair est également présent et ne permet donc pas de garantir la confidentialité des données personnelles. Enfin, l'enchaînement des deux approches et le fait de terminer par les méthodes symboliques permet donc de traiter toutes les informations personnelles nominatives présentes dans le corpus et de n'en laisser aucune en clair.

Enfin, nous rappelons qu'en matière d'anonymisation automatique, on tolèrera toujours un système ayant tendance à sur-anonymiser dès lors qu'il permet d'anonymiser un plus grand nombre d'entités.

Conclusion de la deuxième partie

Dans cette deuxième partie, nous avons rassemblé les différentes expériences que nous avons menées en matières d'anonymisation automatique de documents cliniques.

❧

Nous avons d'abord présenté le guide d'annotation que nous avons constitué pour annoter manuellement le corpus de documents cliniques. Nous avons ensuite détaillé les corpus sur lesquels nous avons réalisé nos expériences et la procédure que nous avons suivie pour produire le corpus de référence. Enfin, nous avons introduit les outils que nous avons utilisés pour préparer les corpus et évaluer les résultats.

❧

Nous avons par la suite détaillé les démarches d'anonymisation reposant sur les méthodes symboliques. Nous avons notamment présenté les trois démarches suivies : un rappel des premières approches de l'anonymisation réalisées en 2002, puis les deux expériences récentes, la première sur la tentative de francisation d'un outil existant, la seconde sur la création *ex nihilo* d'un outil dédié.

❧

Nous avons ensuite présenté les expériences menées à base d'apprentissage statistique et d'hybridation. Nous avons ainsi introduit le protocole expérimental suivi, fondé sur une validation croisée. Après avoir présenté les outils à notre disposition reposant sur le formalisme des CRFs, nous avons exposé les paramètres de configuration que nous avons retenu pour Wapiti, l'outil que nous avons effectivement utilisé. Enfin, nous avons présenté les différentes expériences que nous avons menées.

❧

Enfin, nous avons achevé cette partie par un chapitre portant sur l'évaluation des résultats produits par chacune des méthodes. Nous avons alors discuté des avantages et des inconvénients de chaque type de méthode. Nous avons également établi un bilan du traitement réservé aux informations nominatives personnelles en étudiant quel type de méthode permettait de garantir au maximum l'anonymisation des noms et prénoms des patients.

Conclusion générale

Problématique

Dans le cadre de ce travail de thèse, nous nous sommes intéressé à la problématique de l'anonymisation automatique des données personnelles contenues dans des documents cliniques. Cette problématique consiste à relever et masquer les informations nominatives ou numériques qui permettent d'identifier le patient évoqué dans le document tout en préservant les données cliniques. Nous avons notamment étudié deux principaux types de méthodes : (i) les méthodes dites « symboliques » et (ii) les méthodes à base d'apprentissage statistique, ainsi que la combinaison de ces deux méthodes sous la forme d'une hybridation.

Les méthodes symboliques reposent sur l'utilisation de dictionnaires *(dictionnaire de langue générale ou de termes métiers spécifiques)*, de listes *(listes d'entités nommées : noms, prénoms, villes, etc.)*, de déclencheurs *(des indices permettant de détecter des informations à anonymiser dans leur voisinage : « M., Mme, Dr » pour les noms de personnes, « CHU, clinique » pour les noms d'hôpitaux)* et la définition de patrons syntaxiques. Ces méthodes reposent très largement sur la mobilisation de connaissances d'experts qui sont alors formalisées.

À l'opposé, les approches à base d'apprentissage consistent à déléguer à l'ordinateur le processus de création de ces différentes règles. L'utilisateur fournit alors à la machine une base d'apprentissage constituée d'exemples annotés enrichis d'informations annexes *(caractéristiques de surface, propriétés issues d'outils externes)* et une description détaillée, sous la forme d'un fichier de configuration, de la manière dont les informations annexes ainsi produites doivent être combinées aux exemples annotés.

Il existe plusieurs types d'hybridation des méthodes : (i) l'application successive des deux méthodes *(la seconde prend en entrée le résultat de la première)*, par exemple en définissant des règles de pré- et post-traitements, (ii) l'application parallèle des deux méthodes *(deux anonymisations sont alors produites)* suivie d'une étape de vote entre chaque version produite, et (iii) l'imbrication des deux méthodes, notamment en intégrant dans le processus d'apprentissage des informations provenant de ressources linguistiques *(listes, informations sémantiques, morphologiques, syntaxiques, etc.)*.

Dans ce travail, même si nous avons testé l'enchaînement des deux méthodes, nous nous sommes principalement intéressés au dernier type d'hybridation reposant sur l'imbrication d'informations d'ordre linguistique dans le processus d'apprentissage statistique.

Guide et annotation de corpus

Guide d'annotation. Préalablement à la réalisation de nos différentes expériences, nous avons conçu un guide d'annotation qui donne une vue globale de l'objectif poursuivi et qui détaille les différentes catégories d'informations à anonymiser. Pour chacune des catégories, nous avons fourni une définition et plusieurs exemples relevant de cette catégorie. Les principes généraux d'annotation ont également été présentés *(encadrement de l'information au moyen de balises XML typantes avec préservation de la tokénisation d'origine).*

Annotation de corpus. Sur la base de ce guide d'annotation, nous avons annoté 562 fichiers du corpus de cardiologie provenant du projet Akenaton. Cent fichiers ont fait l'objet d'une double annotation. L'accord inter-annotateurs calculé sur ces cent fichiers renvoie un coefficient κ de 0,8073. Une phase de fusion par adjudication des annotations a été réalisée pour bénéficier d'un corpus de qualité.

Protocole expérimental. Pour la réalisation de nos différentes expériences, nous avons réparti ces 562 fichiers en 62 fichiers pour le corpus de test *(issus de la double annotation, conférant à ce corpus une meilleure qualité)*, 250 fichiers *(dont 38 provenant de la double annotation)* pour le corpus d'apprentissage et 250 fichiers pour le corpus de développement. Le corpus de test a notamment été utilisé pour comparer les approches symboliques et à base d'apprentissage.

D'autre part, nous avons réalisé nos expérimentations à base d'approches statistiques au moyen d'une validation croisée 10-fois, fondée sur les 500 fichiers rassemblant les corpus d'apprentissage et de développement. Cette procédure permet ainsi d'éprouver la robustesse du système sur un plus grand nombre de documents.

Le déclin des méthodes symboliques

Compte tenu de l'avancée des méthodes par apprentissage statistique, nous estimons que les méthodes symboliques ne sont désormais plus viables sur l'ensemble des catégories traitées dans une tâche d'anonymisation automatique. Cette observation est confirmée par les résultats obtenus par les participants de l'édition 2006 du défi i2b2 consacré à l'anonymisation automatique, dans lequel les équipes ayant mobilisé des approches statistiques ont le mieux réussi, avec des écarts élevés compris entre 10 et 18 points de F-mesure *(tableau 1.3)* vis à vis des équipes n'ayant utilisé que des méthodes à base de règles.

Les méthodes symboliques demandent un temps de travail aussi important, si ce n'est plus, que celui nécessaire aux approches par apprentissage. Elles rapportent néanmoins des résultats moins élevés *(un écart de six points de F-mesure a systématiquement été observé sur nos expériences d'anonymisation entre les approches symboliques et celles à base d'apprentissage, voir figure 7.2).*

D'après les expériences que nous avons réalisées, les méthodes symboliques supposent de procéder aux étapes suivantes : (i) étudier le corpus en détail pour identifier les informations à traiter *(quelles catégories d'information)* et la typologie de présentation de ces informations en fonction des caractéristiques structurelles et linguistiques du corpus, (ii) rassembler les ressources *(listes d'entités)* ou compléter les ressources existantes d'après les caractéristiques identifiées, (iii) définir les règles

d'extraction sur la base des observations effectuées en corpus, puis (iv) effectuer de nombreuses et fastidieuses étapes de correction des règles créées, en réalisant plusieurs évaluations successives sur le corpus d'entraînement, pour formaliser au maximum les caractéristiques du corpus.

Au final, nous constatons sur nos données que les méthodes symboliques ont été plus efficaces que les méthodes statistiques uniquement sur les catégories très facilement formalisables telles que les données numériques *(essentiellement les codes postaux et les numéros de téléphone)*. Cependant, sur la catégorie des dates *(qui rassemble des informations uniquement numériques « 15/12/2012 » ou un mélange alphanumérique « 15 décembre 2012 », « du 15 au 18 décembre 2012 », sans compter les scories diverses « juin 2OO5 »)*, les approches par apprentissage ont produit de bien meilleurs résultats, avec un écart constaté de huit points de F-mesure...

L'intérêt des approches par apprentissage

Dans le cadre de nos expériences, pour peu que l'on dispose d'un corpus annoté de qualité, nous avons pu constater que nous obtenons de meilleurs résultats en utilisant des approches à base d'apprentissage *(F-mesure globale de 0,8606 en symbolique et de 0,9211 en apprentissage)*. Nous observons également que nous bénéficions d'un taux de sur-anonymisation plus réduit, conduisant à une meilleure précision *(précision globale de 0,8849 en symbolique contre 0,9646 en apprentissage)*. D'autre part, le nombre d'informations à anonymiser correctement traitées est également largement plus important dans les approches par apprentissage *(rappel de 0,8813)* que dans les méthodes à base de règles *(rappel de 0,8375)*.

L'obtention de bons résultats par ce type d'approche est cependant dépendante de deux facteurs principaux : (i) la qualité des annotations produites sur le corpus servant de base d'apprentissage pour la machine, et (ii) la juste définition des paramètres et caractéristiques des tokens à utiliser pour la construction du modèle. Si la qualité des annotations produites peut et doit être vérifiée tout au long de la phase d'annotation du corpus *(calculs d'accords inter-annotateurs)*, la recherche de la meilleure configuration d'utilisation des caractéristiques produites demeure néanmoins largement empirique et nécessite la réalisation de nombreuses expérimentations.

Le gain apporté par l'hybridation des méthodes

Hybridation par imbrication

Si l'utilisation seule de la forme des tokens ou des caractéristiques de surface permet déjà d'obtenir des résultats valables, leur combinaison et plus encore, leur association avec des ressources linguistiques conduit à une nette amélioration des résultats. Cette combinaison des différents types de caractéristiques produits constitue une hybridation entre l'apprentissage statistique et les ressources linguistiques.

Nous avons mené plusieurs expérimentations en validation croisée reposant sur l'utilisation de caractéristiques de différents types utilisées seules ou combinées. Au niveau global, l'utilisation des seules caractéristiques de surface ne permet pas l'obtention de résultats valables *(F=0,0500)* ; en revanche, l'utilisation des seules ressources externes ou uniquement de la forme des tokens permet une base de travail appréciable *(F-mesures respectivement de 0,8395 et 0,8834)* ; la combinaison des

différents types de caractéristiques permet au modèle créé de gagner en robustesse *(F=0,9010 en combinant formes des tokens et caractéristiques de surface, F=0,9070 en combinant formes des tokens et ressources externes)* ; la combinaison de toutes les caractéristiques produit au final les meilleurs résultats *(F=0,9235)*.

Hybridation par enchaînements

Compte-tenu des performances obtenues par chaque méthodes séparées, nous avons estimé que l'hybridation par enchaînement des méthodes devait reposer en priorité sur les approches à base d'apprentissage, celles-ci ayant obtenu les meilleurs résultats, tout en conservant une approche hybride du point de vue de la construction des caractéristiques. L'application des méthodes symboliques a donc été réalisée sur les sorties produites par l'approche par apprentissage. Cette procédure permet au système à base de règles de compléter les anonymisations en identifiant les informations qui n'auraient pas été traitées par l'apprentissage statistique.

Si l'enchaînement des deux méthodes n'a pas permis d'augmenter la F-mesure globale *(F=0,9235 par Wapiti en validation croisée ; F=0,9217 par l'enchaînement Wapiti/Medina)*, nous notons en revanche que le résultat produit des valeurs de rappel supérieures à celles de la précision sur les catégories des dates, des noms de personne et des noms d'hôpitaux, ce que l'on peut traduire comme étant la manifestation d'une anonymisation plus importante. Autrement dit, l'enchaînement des méthodes permet d'anonymiser correctement plus d'entités, mais elle conduit également à une sur-anonymisation plus importante, que l'on observe par une baisse substancielle dans les valeurs de précision de chaque catégorie.

Traitements des informations personnelles

Si l'anonymisation des comptes rendus cliniques par un appariement avec les informations du SIP de l'hôpital permet de garantir un premier niveau d'anonymisation, nous avons pu constater que 25 % des documents étaient toujours porteurs d'informations personnelles.

L'application d'une deuxième approche se révèle nécessaire pour traiter les informations personnelles laissées en clair à l'issue de cette première étape. Nous avons observé que l'utilisation des approches par apprentissage statistique permet de faire baisser le pourcentage de documents contenant des informations personnelles à 2 % (soit 10 documents sur 500). De manière plus fiable, l'outil que nous avons développé à base de règles permet d'identifier et d'anonymiser toutes les occurrences d'informations personnelles non traitées par la première étape.

Ainsi, si les approches par apprentissage statistique permettent d'obtenir globalement de bien meilleurs résultats que les méthodes symboliques, elles ne permettent pas de garantir un niveau de confidentialité optimal dans la mesure où le modèle construit manquera de robustesse sur certains cas particuliers *(en l'occurrence les prénoms des patients en début de ligne ou entre parenthèses)*. L'enchaînement des méthodes, outre l'apport qualitatif dans l'anonymisation des données, permet d'assurer une plus grande confidentialité des informations personnelles.

Productions et contributions

Productions

Ce travail de thèse s'accompagne de la production de plusieurs éléments.

- En premier lieu, nous avons établi un guide d'annotation dont l'objectif est de constituer un corpus annoté de référence. Ce guide fixe les principes d'annotation, présente des exemples et définit les douze catégories d'information à anonymiser que nous avons retenues.
- Sur la base de ce guide d'annotation, nous avons manuellement annoté un corpus de 562 comptes rendus cliniques en cardiologie. Parmi ces documents, 100 ont fait l'objet d'une double annotation, avec calcul d'accord inter-annotateur *(coefficient κ)*, suivie d'une phase d'adjudication des annotations. Ce corpus de référence a été scindé en trois : un corpus d'entraînement servant de base à l'apprentissage, un corpus de développement pour optimiser la construction du modèle statistique et un corpus de test pour évaluer les résultats des systèmes.
- Enfin, nous avons développé un outil d'aide à l'anonymisation automatique des données personnelles intitulé Medina *(Medical Information Anonymization)*.

Contributions

La première contribution de ce travail de thèse a consisté en l'étude et l'expérimentation des deux grandes méthodes existantes pour anonymiser des informations personnelles dans le domaine clinique, en appliquant les outils que nous avons développés sur le même jeu de données.

La deuxième contribution a concerné l'impact des définitions humaines retenues en matière de catégorisation des informations traitées *(distinction nom/prénom, définition des adresses sans distinguer les composants)* au regard des performances des anonymisations réalisées globalement et pour chacune de ces catégories.

Enfin, la dernière contribution a consisté à situer dans le processus global de l'anonymisation le traitement appliqué aux informations les plus sensibles *(noms et prénoms)* selon la méthode choisie.

Perspectives

Nous prévoyons d'étudier plus en détail les possibilités de combinaison des différentes méthodes, notamment par le biais des procédures de vote. Ce type de procédure permet, à l'intérieur de chaque passage d'un document clinique, soit de conserver l'anonymisation réalisée par l'une ou l'autre des deux méthodes *(choisir l'anonymisation complète et écarter l'anonymisation partielle)*, soit de combiner les deux anonymisations réalisées *(en cas d'anonymisations partielles produites par chacune des méthodes : seulement le nom par une approche, le prénom par l'autre, anonymiser le nom et le prénom comme résultat de cette combinaison)*. Des procédures de vote sont implémentées dans la boîte à outils Weka.

D'autre part, puisque chaque type de méthodes parvient à gérer mieux qu'une autre certains types de catégories, il apparaît pertinent de ne pas chercher à traiter toutes les catégories avec chacune des méthodes, mais uniquement celles qui sont correctement traitées, et d'envisager l'enchaînement des méthodes de manière complémentaire. On réservera ainsi les entités numériques *(codes postaux, numé-*

ros de téléphones, éventuellement les dates) pour les méthodes symboliques qui seront alors appliquées sur les sorties de l'apprentissage sous la forme de règles de post-traitement. Nous considérons que cette combinaison permet de tirer partie du meilleur de chacune des deux méthodes.

Enfin, nous envisageons de tenir compte des poids accordés sur chaque prédiction par les approches à base d'apprentissage statistique, de manière à accorder plus d'importance au rappel qu'à la précision. La définition de seuils spécifiques permettra d'augmenter le rappel au détriment de la précision, conduisant ainsi le système à anonymiser un plus grand nombre d'éléments, parmi lesquels devraient figurer davantage d'informations personnelles.

Bibliographie

[Aberdeen et al., 2010] Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B., et Hirschman, L. (2010). The MITRE identification scrubber toolkit: Design, training, and assessment. *Int J Med Inform*, 79(12):849–59. – Cité pages 73, 85 et 158.

[Adda et al., 1999] Adda, G., Mariani, J., Paroubek, P., Rajman, M., et Lecomte, J. (1999). L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Langues*, 2(2):119–29. – Cité page 162.

[Allauzen et Bonneau-Maynard, 2008] Allauzen, A. et Bonneau-Maynard, H. (2008). Training and evaluation of POS taggers on the French MULTITAG corpus. In *Proc of LREC*, pages 3373–7. – Cité page 162.

[Aramaki et al., 2006] Aramaki, E., Imai, T., Miyo, K., et Ohe, K. (2006). Automatic deidentification by using sentence features and label consistency. In *Proceedings of i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC. – Cité page 80.

[Artstein et Poesio, 2008] Artstein, R. et Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–96. – Cité pages 109, 110, 111 et 134.

[Baude, 2006] Baude, O. (2006). *Corpus Oraux. Guide des Bonnes Pratiques 2006*, édition Presses Universitaires d'Orléans, CNRS. – Cité pages 37 et 47.

[Beckwith et al., 2006] Beckwith, B. A., Mahaadevan, R., Balis, U. J., et Kuo, F. (2006). Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Med Inform Decis Mak*, 6(12). – Cité page 66.

[Benitez et Malin, 2010] Benitez, K. et Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc*, 17(2):169–77. – Cité page 38.

[Bennett et al., 1954] Bennett, E. M., Alpert, R., et Goldstein, A. C. (1954). Communications through limited questioning. *Public Opin Q*, 18(3):303–8. – Cité page 109.

[Benton et al., 2011] Benton, A., Hill, S., Ungar, L., Chung, A., Leonard, C., Freeman, C., et Holmes, J. H. (2011). A system for de-identifying medical message board text. *BMC Bioinformatics*, 12. – Cité page 74.

[Berman, 2003] Berman, J. J. (2003). Concept-match medical data scrubbing. how pathology text can be used in research. *Arch Pathol Lab Med*, 127(6):680–6. – Cité page 66.

[Bindel et Goodman, 2006] Bindel, D. et Goodman, J. (2006). *Principles of Scientific Computing*. New York University. Chapitre 9. Monte Carlo Methods. – Cité page 113.

[Bossy et al., 2012] Bossy, R., Jourde, J., Manine, A.-P., Veber, P., Alphonse, E., van de Guchte, M., Bessières, P., et Nédellec, C. (2012). BioNLP shared task – the bacteria track. *BMC Bioinformatics*, 13(Suppl 11):S3. – Cité page 34.

[Brill, 1992] Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proc of Applied Natural Language Conference*, pages 152–5. – Cité page 162.

[Brown et al., 1992] Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C., et Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–79. – Cité page 161.

[Carletta, 1996] Carletta, J. (1996). Assessing agreement on classification tasks: the Kappa statistics. *Computational Linguistics*, 22(2):249–54. – Cité page 109.

[Chang et Lin, 2011] Chang, C.-C. et Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol*, 2(3):1–27. – Cité page 79.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ Psychol Meas*, 20(1):37–46. – Cité page 110.

[Corbin et Paul, 2000] Corbin, D. et Paul, J. (2000). Aperçus sur la créativité morphologique dans la terminologie de la chimie. *La banque des mots*, 60:51–68. – Cité page 70.

[Culotta et McCallum, 2004] Culotta, A. et McCallum, A. (2004). Confidence estimation for information extraction. In *Proc of HLT*. – Cité page 84.

[Dankar et al., 2012] Dankar, F. K., El Emam, K., Neisa, A., et Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC Med Inform Decis Mak*, 12(66). – Cité page 39.

[Darmoni et al., 2000] Darmoni, S. J., Leroy, J.-P., Baudic, F., Douyère, M., Piot, J., et Thirion, B. (2000). CISMeF: a structured health resource guide. *Methods Inf Med*, 39(1):30–5. – Cité page 21.

[Davies et Fleiss, 1982] Davies, M. et Fleiss, J. L. (1982). Measuring agreement for multinominal data. *Biometrics*, 38(4):1047–51. – Cité page 111.

[Deléger et al., 2013] Deléger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L., et Solti, I. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J Am Med Inform Assoc*, 20(1):84–94. – Cité pages 45, 46, 84, 85 et 88.

[Dress, 2004] Dress, F. (2004). *Probabilités et statistique de A à Z*. Sciences Sup. Dunod, Paris. – Cité page 112.

[Ehrmann, 2008] Ehrmann, M. (2008). *Les entités nommées de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Thèse de Doctorat, Université Paris VII - Denis Diderot. – Cité page 73.

[El Emam et al., 2009] El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., et Bottomley, J. (2009). A globally optimal k-anonymity method for the de-identification of health data. *J Am Med Inform Assoc*, 16(5):670–82. – Cité page 45.

[El Kalam, 2003] El Kalam, A. A. (2003). *Modèles et politiques de sécurité pour les domaines de la santé et des affaires sociales*. Thèse de Doctorat, Université de Toulouse. – Cité page 40.

[Farfor, 1976] Farfor, J. A. (1976). *Cours élémentaire de rédaction médicale*. Cahiers médicaux, Lyon. Traduit de l'anglais par J Feisthauer. – Cité page 35.

[Ferrández et al., 2012] Ferrández, O., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H., et Meystre, S. M. (2012). Generalizability and comparison of automatic clinical text de-identification methods and resources. In *AMIA Annu Symp Proc*, Chicago, IL. – Cité pages 62, 84, 92 et 103.

[Fielstein et al., 2004] Fielstein, E. M., Brown, S. H., et Speroff, T. (2004). Algorithmic de-identification of VA medical exam text for HIPAA privacy compliance: Preliminary findings. In *Proc of MedInfo*, San Francisco, CA. – Cité page 66.

[Fleischman et Hovy, 2002] Fleischman, M. et Hovy, E. (2002). Fine grained classification of named entities. In *Proc of COLING*, pages 1–7, Taipei, Taïwan. – Cité page 73.

[Fleiss, 1971] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol Bull*, 76(5):378–82. – Cité page 111.

[Fort, 2012] Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Thèse de Doctorat, Université Paris 13. – Cité page 125.

[Friedlin et McDonald, 2008] Friedlin, F. J. et McDonald, C. J. (2008). A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc*, 15(5):601–10. – Cité pages 66 et 68.

[Friedman et al., 1994] Friedman, C., Alderson, P. O., Austin, J. H. M., Cimino, J. J., et Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, 1(2):161–74. – Cité page 90.

[Galibert et al., 2011] Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., et Quintard, L. (2011). Structured and extended named entity evaluation in automatic speech transcriptions. In *Proc of IJCNLP*, Chiang Mai, Thailand. – Cité pages 105 et 136.

[Gardner et Xiong, 2008] Gardner, J. et Xiong, L. (2008). HIDE: An integrated system for health information de-identification. In *Proc of International Symposium on Computer-Based Medical Systems*, pages 254–9. – Cité page 81.

[Gardner et Xiong, 2009] Gardner, J. et Xiong, L. (2009). An integrated framework for de-identifying unstructured medical data. *Data Knowl Eng*, 68:1441–51. – Cité pages 52 et 91.

[Gicquel et al., 2012] Gicquel, Q., Proux, D., Marchal, P., Hagège, C., Berrouane, Y., Darmoni, S. J., Pereira, S., Segond, F., et Metzger, M.-H. (2012). Evaluation d'un outil d'aide à l'anonymisation des documents médicaux basé sur le traitement automatique du langage naturel. *Informatique et Santé*, 1:165–76. – Cité pages 50 et 74.

[Golle, 2006] Golle, P. (2006). Revisiting the uniqueness of simple demographics in the us population. In *Proc of WPES*. – Cité page 38.

[Grabar, 2004] Grabar, N. (2004). *Terminologie médicale et morphologie. Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique*. Thèse de Doctorat, Université Paris 6. – Cité page 34.

[Green, 1997] Green, A. M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proc of the 22nd Annual SAS Users Group International Conference*. – Cité page 111.

[Grishman et Sundheim, 1996] Grishman, R. et Sundheim, B. (1996). Message understanding conference - 6: A brief history. In *Proc of COLING*, pages 466–71, Copenhague, Danemark. – Cité page 72.

[Grouin, 2002] Grouin, C. (2002). *Chaîne de traitements pour la constitution automatique de corpus : application au domaine médical pour le projet corpus CLEF*. Mémoire de fin d'études, DESS ingénierie multilingue, INaLCO, Paris. CHU de la Pitié-Salpêtrière, AP-HP/DSI/STIM. – Cité page 140.

[Grouin et al., 2009a] Grouin, C., Rosier, A., Dameron, O., et Zweigenbaum, P. (2009a). Testing tactics to localize de-identification. *Stud Health Technol Inform*, 150:735–9. – Cité page 143.

[Grouin et al., 2009b] Grouin, C., Rosier, A., Dameron, O., et Zweigenbaum, P. (2009b). Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. In Fieschi, M., Staccini, P., Bouhaddou, O., et Levis, C. (éditeurs), *Risques, technologies de l'information pour les pratiques médicales*, chapitre XVII, pages 23–34. Springer-Verlag, Nice, France. – Cité page 54.

[Grouin et al., 2011] Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., et Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proc of Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, OR. Association for Computational Linguistics. – Cité pages 136 et 137.

[Grouin et Zweigenbaum, 2011] Grouin, C. et Zweigenbaum, P. (2011). Une approche à plusieurs étapes pour anonymiser des documents médicaux. *RSTI–RIA*, 25:525–49. – Cité page 166.

[Guo et al., 2006] Guo, Y., Gaizauskas, R., Roberts, I., Demetriou, G., et Hepple, M. (2006). Identifying personal health information using support vector machines. In *Proc of i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC. – Cité pages 52 et 79.

[Gupta et al., 2004] Gupta, D., Saul, M., et Gilbertson, J. (2004). Evaluation of a deidentification (De-ID) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol*, 121(2):176–86. – Cité pages 66 et 67.

[Hara, 2006] Hara, K. (2006). Applying a SVM based chunker and a text classifier to de-id challenge. In *Proc of i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC. – Cité pages 77 et 78.

[Harris, 1968] Harris, Z. S. (1968). *Mathematical structures of language*. Wiley Interscience, New York, NY. – Cité page 34.

[Harris, 1985] Harris, Z. S. (1985). Distributional structure. In Katz, J. (éditeur), *The Philosophy of Linguistics*, pages 26–47. Oxford University Press. – Cité page 87.

[Hripcsak et Rothschild, 2005] Hripcsak, G. et Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc*, 12(3):296–8. – Cité page 110.

[Kim et al., 2003] Kim, J. D., Ohta, T., Tateisi, Y., et Tsujii, J. (2003). GENIA corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1):180–2. – Cité page 33.

[Kittredge et Lehrberger, 1982] Kittredge, R. et Lehrberger, J. (éditeurs) (1982). *Sublanguage Studies of language in restricted semantic domains*. Walter de Gruyter, Berlin, New York, NY. – Cité page 34.

[Krippendorff, 1980] Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage Publications. – Cité page 111.

[Kudo et al., 2004] Kudo, T., Yamamoto, K., et Matsumoto, Y. (2004). Applying conditional random fields to japanese morphological analysis. *Proc of EMNLP*, pages 230–7. – Cité page 157.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., et Pereira, F. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*. – Cité page 80.

[Landi et Rao, 2003] Landi, W. et Rao, R. B. (2003). Secure de-identification and re-identification. In *AMIA Annu Symp Proc*, page 905. – Cité page 37.

[Landis et Koch, 1977] Landis, J. R. et Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–74. – Cité page 111.

[Lavergne et al., 2010] Lavergne, T., Cappé, O., et Yvon, F. (2010). Practical very large scale CRFs. *Proc of ACL*, pages 504–13. – Cité page 158.

[Le Barbanchon, 2012] Le Barbanchon, T. (2012). *Essays on Labor Market Policies Evaluation*. Thèse de Doctorat, École Polytechnique. – Cité page 39.

[Liang, 2005] Liang, P. (2005). Semi-supervised learning for natural language. Master's thesis, MIT. – Cité page 161.

[Lindberg et al., 1993] Lindberg, D. A., Humphreys, B. L., et McRay, A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, 32(4):281–91. – Cité page 66.

[Lip, 2013] Lip, G. Y. (2013). Using the CHA2DS2-VASc score for stroke risk stratification in atrial fibrillation: a clinical perspective. *Expert Rev Cardiovasc Ther*, 11(3):259–62. – Cité page 129.

[Loukides et al., 2010] Loukides, G., Gkoulalas-Divanis, A., et Malin, B. (2010). Anonymization of electronic medical records for validating genome-wide association studies. *Proc Natl Acad Sci USA*, 107(17):7898–903. – Cité page 55.

[Machanavajjhala et al., 2006] Machanavajjhala, A., Gehrke, J., Kifer, D., et Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *Proc of ICDE*. – Cité pages 42 et 92.

[Makhoul et al., 1999] Makhoul, J., Kubala, F., Schwartz, R., et Weischedel, R. (1999). Performance measures for information extraction. In *Proc. of DARPA Broadcast News Workshop*, pages 249–52. – Cité page 102.

[Malin et al., 2011] Malin, B., Benitez, K., et Masys, D. (2011). Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA privacy rule. *J Am Med Inform Assoc*, 18(1):3–10. – Cité page 46.

[Manning et Schütze, 2000] Manning, C. D. et Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts. – Cité pages 78, 97, 98 et 113.

[Mayer et al., 2009] Mayer, J., Shen, S., South, B. R., Meystre, S., Friedlin, F. J., Ray, W. R., et Samore, M. (2009). Inductive creation of an annotation schema and a reference standard for de-identification of va electronic clinical notes. In *AMIA Annu Symp Proc*, pages 416–20. – Cité page 125.

[McCallum, 2002] McCallum, A. (2002). *MALLET: A Machine Learning for Language Toolkit*. – Cité pages 84 et 157.

[McCallum, 2003] McCallum, A. (2003). Efficiently inducing features of Conditional Random Fields. In *Proc of Conf Uncertain Artif Intell*, pages 403–10, Acapulco, Mexico. Morgan Kaufmann. – Cité pages 80 et 164.

[McDonald, 1993] McDonald, D. D. (1993). Internal and external evidence in the identification and semantic categorization of proper names. In Boguraev, B. et Pustejovsky, J. (éditeurs), *Corpus Processing for Lexical Acquisition*, pages 61–76, Cambridge, MA. MIT Press. – Cité page 86.

[McGraw, 2013] McGraw, D. (2013). Building public trust in uses of health insurance portability and accountability act de-identified data. *J Am Med Inform Assoc*, 20(1):29–34. – Cité page 45.

[Metropolis et Ulam, 1949] Metropolis, N. et Ulam, S. (1949). The Monte Carlo Method. *J Am Stat Assoc*, 44(247):335–41. – Cité page 113.

[Meystre et al., 2010] Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., et Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol*, 10(70). – Cité pages 37, 65 et 86.

[Minkov et al., 2006] Minkov, E., Wang, R. C., Tomasic, A., et Cohen, W. W. (2006). NER systems that suit user's preferences: Adjusting the recall-precision trade-off for entity extraction. In *Proc of HLTC*, pages 93–6. – Cité page 84.

[Morehead, 2001] Morehead, P. D. (2001). *The New American Roget's College Thesaurus in Dictionary Form*. Penguin. – Cité page 36.

[Morrison et al., 2009a] Morrison, F. P., Li, L., Lai, A. M., et Hripcsak, G. (2009a). Repurposing the clinical record: Can an existing natural language processing system de-identify clinical notes? *J Am Med Inform Assoc*, 16(1):37–9. Technical Brief. – Cité page 66.

[Morrison et al., 2009b] Morrison, F. P., Sengupta, S., et Hripcsak, G. (2009b). Using a pipeline to improve de-identification performance. In *AMIA Annu Symp Proc*, pages 447–51. – Cité page 90.

[Namer, 2000] Namer, F. (2000). Flemm : un analyseur flexionnel du français à base de règles. *TAL*, 41(2):523–47. – Cité page 162.

[Neamatullah, 2006] Neamatullah, I. (2006). *De-Identification of Free-Text Medical Records*. MIT, édition 1.1. – Cité pages 54, 67 et 143.

[Neamatullah et al., 2008] Neamatullah, I., Douglass, M. M., Lehman, L.-W. H., Reisner, A., Villaroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., et Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*, 8(32). – Cité pages 54, 67 et 143.

[Olesen et al., 2012] Olesen, J. B., Torp-Pedersen, C., Hansen, M. L., et Lip, G. Y. (2012). The value of the CHA2DS2-VASc score for refining stroke risk stratification in patients with atrial fibrillation with a CHADS2 score 0-1: A nationwide cohort study. *Thromb Haemost*, 107(6):1172–9. – Cité page 129.

[Pelletier et al., 2004] Pelletier, F., Plamondon, L., et Lapalme, G. (2004). L'assistant d'anonymisation NOME. In *Journées Internet pour le Droit*, Paris. – Cité page 74.

[Pestian et al., 2007] Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., Bretonnel-Cohen, K., et Duch, W. (2007). A shared task involving multi-label classification of clinical free text. In *Proc of BioNLP*, pages 97–104, Prague. Association for Computational Linguistics. – Cité pages 33, 34 et 100.

[Pestian et al., 2012] Pestian, J. P., Matykiewicz, P., et Linn-Gust, M. (2012). What's in a note: construction of a suicide note corpus. *Biomed Inform Insights*, 5:1–6. – Cité page 41.

[Plamondon et al., 2004] Plamondon, L., Lapalme, G., et Pelletier, F. (2004). Anonymisation de décisions de justice. In *Actes de TALN*. – Cité page 73.

[Péchoin, 1999] Péchoin, D. (1999). *Thésaurus*. Larousse. – Cité page 36.

[Quantin et al., 2005] Quantin, C., Gouyon, B., Allaert, F.-A., et Cohen, O. (2005). Proposition d'un identifiant individuel à composante familiale pour l'identification européenne du patient dans le secteur de la santé. In *Actes des JFIM*, Lille. – Cité page 130.

[Ratinov et Roth, 2009] Ratinov, L. et Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proc of CoNLL*. – Cité page 82.

[Reffay et Teutsch, 2007] Reffay, C. et Teutsch, P. (2007). Anonymisation de corpus réutilisables. Masquer l'identité sans altérer l'analyse des intéractions. In *Environnements Informatiques pour l'Apprentissage Humain*, Lausanne, Suisse. – Cité pages 37, 47 et 106.

[Rey-Debove et Rey, 1993] Rey-Debove, J. et Rey, A. (1993). *Le nouveau Petit Robert*. Le Petit Robert, Paris. – Cité page 35.

[Riedmiller et Braun, 1993] Riedmiller, M. et Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Int Conf Neural Netw*. – Cité page 158.

[Rosset, 2010] Rosset, S. (2010). Methodology for guidelines. In *Quaero Best Practices Annotation Workshop*, Paris. – Cité pages 124 et 125.

[Ruch et al., 2000] Ruch, P., Baud, R. H., Rassinoux, A.-M., Bouillon, P., et Robert, G. (2000). Medical document anonymization with a semantic lexicon. In *AMIA Annu Symp Proc*, pages 729–33. – Cité pages 68, 141 et 177.

[Sapir, 1921] Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. Harcourt, Brace and Company, New York. – Cité page 34.

[Sapir, 1929] Sapir, E. (1929). The status of linguistics as a science. *Language*, 5. – Cité page 34.

[Savova et al., 2010] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., et Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–13. – Cité page 92.

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language*. – Cité page 162.

[Scott, 1955] Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opin Q*, 19(3):321–5. – Cité page 109.

[Sekine, 2004] Sekine, S. (2004). Definition, dictionaries and tagger of extended named entity hierarchy. In *Proc of LREC*. – Cité page 72.

[Sekine et Ranchhod, 2009] Sekine, S. et Ranchhod, E. (éditeurs) (2009). *Named Entities*. John Benjamins Publishing. – Cité page 72.

[Sibanda et Uzuner, 2006] Sibanda, T. C. et Uzuner, O. (2006). Role of local context in automatic deidentification of ungrammatical, fragmented text. In *Proc of HLTC*, pages 65–73. ACL. – Cité page 67.

[Sutton et McCallum, 2006] Sutton, C. et McCallum, A. (2006). An introduction to conditional random fields for relational learning. In Getoor, L. et Taskar, B. (éditeurs), *Introduction to Statistical Relational Learning*. MIT Press. – Cité page 80.

[Sweeney, 1996] Sweeney, L. (1996). Replacing personally-identifying information in medical records, the scrub system. In *AMIA Annu Fall Symp Proc*, pages 333–7, Washington, DC. – Cité pages 41 et 68.

[Sweeney, 2000] Sweeney, L. (2000). Uniqueness of simple demographics in the u.s. population. Data privacy working paper 3, Carnegie Mellon University, Pittsburgh. – Cité pages 38 et 41.

[Sweeney, 2002] Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzz*, 10(5):557–70. – Cité pages 41 et 92.

[Szarvas et al., 2007] Szarvas, G., Farkas, R., et Busa-Fekete, R. (2007). State-of-the-art anonymization of medical records using an iterative machine learning framework. *J Am Med Inform Assoc*, 14(5):574–80. – Cité pages 52 et 78.

[Taira et al., 2002] Taira, R. K., Bui, A. A., et Kangarloo, H. (2002). Identification of patient name references within medical documents using semantic selectional restrictions. In *AMIA Annu Symp Proc*, pages 757–61, Washington, DC. – Cité page 67.

[Tellier, 2012] Tellier, I. (2012). SEM guide d'annotation en chunks. Technical report, LaTTICe, CNRS. – Cité page 163.

[Tellier et al., 2012] Tellier, I., Duchier, D., et Eshkol, I. (2012). Apprentissage automatique d'un chunker pour le français. In *Actes de TALN*. – Cité page 163.

[Thomas et al., 2002] Thomas, S. M., Mamlin, B., Schadow, G., et McDonald, C. J. (2002). A successful technique for removing names in pathology reports using an augmented search and replace method. In *AMIA Annu Symp Proc*, pages 777–81, Washington, DC. – Cité page 67.

[Tu et al., 2010] Tu, K., Klein-Geltink, J., Mitiku, T. F., Mihai, C., et Martin, J. (2010). De-identification of primary care electronic medical records free-text data in Ontario, Canada. *BMC Med Inform Decis Mak*, 10(35). – Cité pages 55 et 67.

[Uzuner et al., 2012] Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J. P., et South, B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc*, 19(5):786–91. – Cité page 34.

[Uzuner et al., 2007] Uzuner, O., Luo, Y., et Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, 14(5):550–63. – Cité pages 43, 49, 51, 52, 75 et 87.

[Uzuner et al., 2008] Uzuner, O., Sibanda, T. C., Luo, Y., et Szolovits, P. (2008). A de-identifier for medical discharge summaries. *Artif Intell Med*, 42(1):13–35. – Cité pages 43, 79 et 86.

[Uzuner et al., 2010] Uzuner, O., Solti, I., et Cadag, E. (2010). Extracting medication information from clinical text. *J Am Med Inform Assoc*, 17(5):514–8. – Cité page 33.

[Uzuner et al., 2011] Uzuner, O., South, B. R., Shen, S., et DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–6. – Cité page 34.

[Uzuner et al., 2006] Uzuner, O., Szolovits, P., et Kohane, I. (2006). i2b2 workshop on natural language processing challenges for clinical records. In *AMIA Annu Symp Proc*. – Cité page 51.

[Vapnik, 1995] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag. – Cité page 78.

[Velupillai et al., 2009] Velupillai, S., Dalianis, H., Hassel, M., et Nilsson, G. H. (2009). Developing a standard for de-identifying electronic patient records written in swedish: Precision, recall and f-measure in a manual and computerized annotation trial. *Int J Med Inform*, 78(12):19–26. – Cité page 54.

[Wellner, 2009] Wellner, B. (2009). *Sequence Models and Ranking Methods for Discourse Parsing*. Thèse de Doctorat, Brandeis University. – Cité pages 73 et 85.

[Wellner et al., 2007] Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J., et Hirschman, L. (2007). Rapidly retargetable approaches to de-identification in medical records. *J Am Med Inform Assoc*, 14(5):564–73. – Cité pages 52, 80 et 84.

[Whorf, 1940] Whorf, B. L. (1940). Science and linguistics. *Technol Rev*, 42(6):229–31. – Cité page 34.

[Whorf, 1956] Whorf, B. L. (1956). The relation of habitual thought and behavior to languages. In Carroll, J. B. (éditeur), *Language, thought and reality*, pages 238–45. MIT Press, Cambridge, MA. – Cité page 34.

[Wisniewski, 2007] Wisniewski, G. (2007). *Apprentissage dans les espaces structurés. Application à l'étiquetage de séquences et à la transformation automatique de documents*. Thèse de Doctorat, UPMC. – Cité pages 75 et 78.

[Zweigenbaum, 2008] Zweigenbaum, P. (2008). Natural language processing in the medical and biomedical domains : A parallel perspective. In Rebholz-Schuhmann, D., Salakoski, T., et Pyysalo, S. (éditeurs), *Proceedings of the Third International Symposium for Semantic Mining in Biomedicine*, pages 3–4, Turku. Centre for Computer Science. Communication invitée. – Cité page 49.

Troisième partie

Annexes

Annexe A

Guide d'annotation « Anonymisation de documents cliniques »

Sommaire

A.1 Introduction

Ce guide reprend les principales lignes directrices qui ont conduit, d'une part à réaliser le corpus de référence anonymisé, et d'autre part à configurer les outils d'anonymisation des comptes rendus médicaux dans le cadre du projet Akenaton [1]. Ce guide tient compte des évolutions qui ont été retenues après la fin du projet, pour poursuivre cette thématique d'anonymisation.

On parle généralement d'anonymisation en français et de de-identification en anglais. Il n'est pas sûr qu'une différence autre que linguistique existe entre ces deux termes. Dans les deux cas, l'objectif recherché consiste à ne pas pouvoir remonter jusqu'à une personne à partir d'un document clinique.

Après avoir brièvement présenté les démarches existantes en la matière, nous présentons les principes généraux retenus dans le cadre de ce travail, puis les catégories d'entités ayant fait l'objet d'une anonymisation.

1. Automated Knowledge from medical records iN Association with a Telecardiology Observation Network, Numéro de financement ANR-07-TecSan-001-06.

A.2 Éléments à anonymiser

En France, il n'existe pas de liste précise indiquant les éléments devant faire l'objet d'une anonymisation, en particulier dans le cas des comptes rendus médicaux. La CNIL incite toutefois à procéder à des travaux de chiffrement des données dans le cas d'anonymisation voulue irréversible. L'anonymisation de données n'entre pas complètement dans ce cas de figure, puisque seules des portions de texte à l'intérieur d'un compte rendu médical doivent faire l'objet d'une telle anonymisation.

Aux Etats-Unis, 18 identifiants (*Personnal Health Information*) ont été définis dans le cadre du *Health Insurance Portability and Accountability Act* [2] :

1. Noms.
2. Toute subdivision géographique plus petite qu'un Etat.
3. Tout élément de date (sauf les années) et les âges (si supérieurs à 90 ans).
4. Numéro de téléphone.
5. Numéro de télécopie.
6. Adresses de messagerie électronique.
7. Numéro de sécurité sociale.
8. Numéro de dossier médical.
9. Numéro de mutuelle.
10. Numéro de compte.
11. Numéro de carte d'identité (permis de conduire).
12. Numéro d'immatriculation et identifiants du véhicule.
13. Références d'appareillage et numéro de série.
14. Adresses Internet.
15. Adresses IP.
16. Identifiants biométriques.
17. Photographies.
18. N'importe quel autre identifiant ou caractéristique unique permettant l'identification (cicatrice, tatouage, etc.).

A.3 Principes

Nous avons retenu les principes suivants lors des phases d'annotation du corpus de référence :

- L'anonymisation est effectuée au moyen de balises XML typantes (chaque balise renvoie au type sémantique de l'entité anonymisée).
- Le type d'annotation retenue consiste à encadrer l'entité à anonymiser par des balises typantes ouvrante et fermante.
- Toute anonymisation doit être effectuée avec un empan maximal (si plusieurs entités successives relèvent d'une seule catégorie – deux prénoms qui se suivent, un numéro de rue suivi du type et du nom de voie –, alors elles seront rassemblées sous une seule balise), sauf dans le cas défini au point suivant.

2. HIPAA : http://www.hhs.gov/ocr/privacy/

– Le formatage du texte d'origine doit être préservé. En conséquence, si deux entités relevant d'une même catégorie se trouvent, pour la première en fin de ligne et pour la seconde en début de ligne suivante, toutes deux seront anonymisées séparément.
– Les signes de ponctuation collant une entité et non constitutifs de l'entité, ne sont pas inclus dans l'empan de l'anonymisation (une entité suivie d'une virgule). Aucune espace ne doit être ajoutée autour de la ponctuation. En revanche, la ponctuation à l'intérieur d'une entité ou constitutive de l'entité (le point d'abréviation dans l'initiale d'un prénom, le séparateur entre jour, mois et année dans une date, le tiret dans un nom ou prénom composé) est intégrée dans la balise d'anonymisation.

A.4 Catégories

La liste des éléments traités lors des travaux d'anonymisation dans le projet Akenaton ont été les suivants. Ils sont inspirés des catégories définies par le HIPAA et résultent d'une analyse des documents composant le corpus.

Certains types d'entité, initialement envisagés comme devant faire l'objet d'une anonymisation, ne l'ont finalement pas été. Le maintien de ces catégories n'induit pas nécessairement une réidentification mais il permet un meilleur accès au sens dans le processus futur d'extraction d'informations.

A.4.1 Personnes

Nom. Le nom de famille correspondant à une personne *(patient, membre de la famille du patient, médecin traitant, chirurgien, interne, personnel de l'hôpital, etc.)*. Les noms de famille donnés à des salles opératoires ou des hôpitaux n'entrent pas dans cette catégorie. Le déclencheur *(M., Mme, Melle)* n'est pas intégré dans la portée de l'anonymisation.

– *Professeur H. POEHEI,*→ *Professeur H. <nom>POEHEI</nom>,*
– *Hervé LE BON,*→ *Hervé <nom>LE BON</nom>,*
– *Madame Pimentel Gisèle,*→ *Madame <nom>Pimentel</nom> Gisèle,*

Prénom. Le prénom d'une personne (mêmes catégories de personnes et mêmes exceptions que ci-dessus).

– *Professeur H. POEHEI,*→ *Professeur <prenom>H.</prenom> POEHEI,*
– *Monsieur Marc-André Senten,*→ *Monsieur <prenom>Marc-André</prenom> Senten,*
– *Madame Pimentel Gisèle,*→ *Madame Pimentel <prenom>Gisèle</prenom>,*
– *Monsieur John F. Kennedy*→ *Monsieur <prenom>John F.</prenom> Kennedy*

Remarque : Dans le premier exemple, le point est constitutif de l'entité anonymisée (point d'abréviation de l'initiale du prénom), il est intégré dans la portée de l'anonymisation ; dans le troisième exemple, la virgule qui suit le prénom n'est pas constitutive de l'entité, elle n'est donc pas intégrée dans la portion anonymisée.

Remarque : Les initiales de second prénom doivent être anonymisées dans la même portion que celle du premier prénom, quatrième exemple.

A.4.2 Lieux

Hôpital. Le nom de l'hôpital *(au sens large : clinique, maison de jour, maison de retraite, maison de repos)*, le nom d'une salle d'opération dans l'hôpital, etc. Le déclencheur *(clinique, hôpital, salle, unité, etc.)* est intégré dans la portion anonymisée.

- *à l'Hôpital Cochin* → *à <hopital>l'Hôpital Cochin</hopital>*
- *Interne, unité Ch. Landron* → *Interne, <hopital>unité Ch. Landron</hopital>*
- *Hôpital de Nantes* → *<hopital>Hôpital de Nantes</hopital>*

Remarque : L'article précédent le déclencheur (généralement « l' »), parce qu'il est collé au déclencheur, doit être intégré dans la portion anonymisée (troisième exemple).

Adresse. L'adresse postale *(numéro, type et nom de voirie, boîte postale, numéro d'escalier et de bâtiment)* du patient, du médecin traitant, de l'hôpital.

- *9, rue de Saint-Malo 35000 Rennes* → *<adresse>9, rue de Saint-Malo</adresse> 35000 Rennes*
- *151 boulevard de l'Hôpital, Bâtiment H, 75013 Paris* → *<adresse>151 boulevard de l'Hôpital, Bâtiment H</adresse>, 75013 Paris*

Code postal. Complémentaire de l'adresse postale.

- *9, rue de Saint-Malo 35000 Rennes* → *9, rue de Saint-Malo <codepostal>35000</codepostal> Rennes*

Ville. Nom de ville figurant dans une adresse postale. En revanche, un nom de famille correspondant à un nom de ville doit être annoté comme un nom.

- *9, rue de Saint-Malo 35000 Rennes* → *9, rue de Saint-Malo 35000 <ville>Rennes</ville>*
- *Monsieur Jacques d'Auvers* → *Monsieur Jacques <nom>d'Auvers</nom>*

A.4.3 Données numériques

Sont regroupées ici les données (essentiellement) numériques autres que le code postal et le numéro de voie dans les adresses postales. Les informations numériques relatives aux appareillages figurent dans la subsection suivante.

Âge. Uniquement les âges si supérieur à 90 ans.

- *70 ans* → *70 ans*
- *90 ans* → *<age> 90 ans </age>*

Date. N'importe quelle date *(date de naissance, date d'entrée ou de sortie, date d'opération chirurgicale, etc.)*, y compris les années. Les dates relatives *(depuis 20 ans)*, les durées *(pendant 3 mois)* ne sont pas anonymisées (quatrième et cinquième exemple).

- *né le 1er mars 2004* → *né le <date>1er mars 2004</date>*
- *(29/09/57)* → *(<date>29/09/57</date>)*
- *jusqu'en 1998* → *jusqu'en <date>1998</date>*
- *depuis 20 ans* → *depuis 20 ans*
- *pendant 3 mois* → *pendant 3 mois*

Les intervalles de dates sont regroupés en une seule portion *(en réduisant au maximum la présence de mots outils)* tandis que les conjonctions et disjonctions de dates seront séparés en deux portions.

- *du 15 au 18 mars* → *du <date>15 au 18 mars</date>*
- *les 15 et 16 avril* → *les <date>15</date> et <date>16 avril</date>*
- *en juin et en septembre 2005* → *en <date>juin</date> et en <date>septembre 2005</date>*
- *le 5 septembre 2004 et le 21 janvier 2005* → *le <date>5 septembre 2004</date> et le <date>21 janvier 2005</date>*

Numéro. N'importe quel identifiant numérique unique relatif au patient ou à l'équipe chirurgicale. Non trouvé dans le corpus.

Sécurité sociale. Numéro de sécurité sociale, de mutuelle, etc., propres au patient.

- *1 87 08 87 227 035* → *<numero_ss>1 87 08 87 227 035</numero_ss>*

Téléphone. Numéro de téléphone, de télécopie, extension téléphonique, numéro de poste, etc. Si deux extensions de numéro sont possibles, séparées par une conjonction, l'ensemble est anonymisé sous une seule balise.

- *Prise de RDV 02.60.10.25.00 ou 01* → *Prise de RDV <telephone>02.60.10.25.00 ou 01</telephone>*
- *nous recontacter au 02.60.10.25.00-01* → *nous recontacter au <telephone>02.60.10.25.00-01</telephone>*

A.4.4 Données médicales

Info. Marque et type d'appareillage *(pacemaker, défibrillateur)*, numéro de série, etc.

- *pacemaker double chambre de type ELA MEDICAL* → *pacemaker double chambre de type <info>ELA MEDICAL</info>*
- *pacemaker de marque ELA MEDICAL, de type BRIO DR 212.* → *pacemaker de marque <info>ELA MEDICAL</info>, de type <info>BRIO DR 212</info>.*
- *une sonde de défibrillation de marque Guidant n° de série : SN0123456789* → *une sonde de défibrillation de marque <info>Guidant</info> n° de série : <info>SN0123456789</info>*

Annexe B

Manuel d'utilisation de Médina

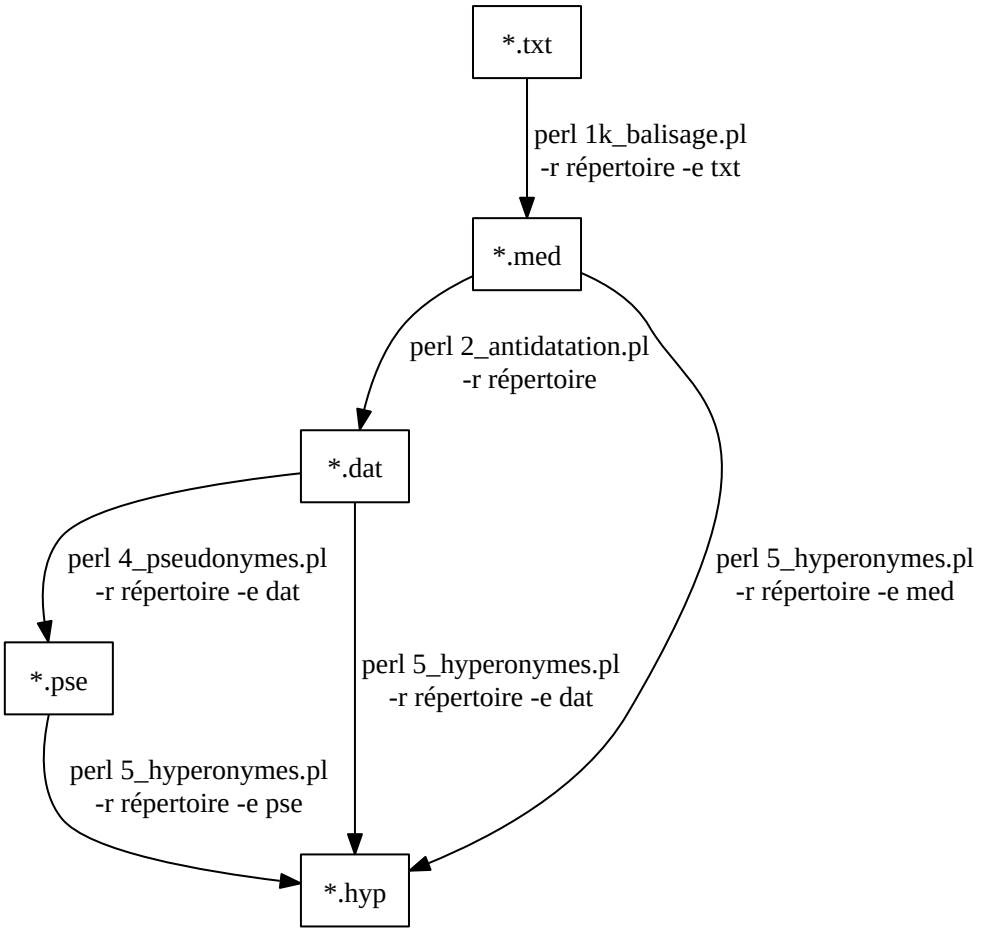
Sommaire

B.1 Présentation

Medina est un outil d'anonymisation des données personnelles présentes dans des documents textuels. Cet outil a été développé pour traiter des comptes rendus cliniques en cardiologie. L'outil se compose de plusieurs scripts, un premier permettant le balisage des informations à anonymiser, suivi de scripts de post-traitements pour procéder à l'anonymisation. L'outil a été développé entre 2008 et 2012 dans le cadre du projet Akenaton pour anonymiser des comptes rendus médicaux en cardiologie. Si certains types d'informations personnelles sont transverses aux différentes disciplines médicales *(nom, prénom, adresse, téléphone, numéro de sécurité sociale...)*, d'autres sont spécifiques *(marques de défibrillateurs en cardiologie, références des dents en stomatologie, etc.)*. Les types de données traitées s'inspirent de la législation américaine HIPAA. Medina ne réalise pas la distinction entre médecin et patient mais conserve la distinction entre nom et prénom. Il est possible d'anonymiser par suppression des informations (hyperonymes) ou par remplacements (antidatation, pseudonymes).

B.2 Lancement rapide

Deux options sont utiles :
-r : répertoire contenant les fichiers (obligatoire) ;
-e : extension des fichiers en entrée (inutile pour le script des dates).



B.2.1 Balisage des informations

```
perl 1k_balisage.pl –r <répertoire> –e <extension des fichiers>
```

B.2.2 Post-traitements

Modification des dates (antidatation comprise entre 1 et 4 ans) ; fichier.med → fichier.dat :

```
perl 2_antidatation.pl –r <répertoire>
```

Remplacement des noms et prénoms par un pseudonyme ; fichier.dat → fichier.pse :
```
perl 4_pseudonymes.pl –r <répertoire> –e <extension des
fichiers>
```

Remplacement des données personnelles par un hyperonyme ; fichier.pse → fichier.hyp :
```
perl 5_hyperonymes.pl –r <répertoire> –e <extension des
fichiers>
```

B.3 Utilisation détaillée

B.3.1 Architecture globale de l'outil

L'outil se compose des éléments suivants :
- un fichier de configuration : config ;
- un répertoire constitué des ressources linguistiques : data *(dictionnaire de mots communs, liste noire de mots ne devant pas être anonymisés, listes de noms de médecins, de noms d'hôpitaux, de noms de famille, de pays, de prénoms et de villes)* ;
- un script pour baliser les informations personnelles : 1k_balisage.pl
- des scripts annexes de post-traitements :
  - 2_antidatation.pl : remplace les dates précédemment identifiées par d'autres dates en conservant l'écart temporel entre chaque date à l'intérieur d'un même document ;
  - 4_pseudonymes.pl : remplace les occurrences de noms et prénoms par des pseudonymes (les dix noms les plus portés en France et dix prénoms mixtes) ;
  - 5_hyperonymes.pl : remplace les données précédemment identifiées par un hyperonyme (conservation des dates si script appliqué sur des fichiers *.dat ; conservation des noms et prénoms si appliqué sur des fichiers *.pse).

B.3.2 Configuration et lancement

Balisage des informations personnelles

L'outil d'anonymisation repose sur une première phase de repérage des informations à anonymiser. À l'issue de ce repérage, les informations seront encadrées de balises XML typant l'information identifiée. Le script produit des fichiers d'extension « *.med » dans le répertoire des documents.

1. Ouvrir le fichier de configuration avec un éditeur de texte et modifier les différents champs selon les besoins :
   - indiquer les informations qui doivent être anonymisées face à chaque catégorie *(adresses, âges, codes postaux, dates, hôpitaux, médicaments, mesures, noms, prénoms, numéro de sécurité sociale, référence des stimulateurs cardiaques, téléphones, unités hospitalières, villes)* ;
   - indiquer les listes de ressources linguistiques à utiliser ;
   - compléter les listes de déclencheurs ;
   - indiquer l'âge minimum au-delà duquel l'anonymisation de l'âge des patients est requise (la législation américaine HIPAA impose d'anonymiser les âges au-delà de 90 ans) ;
   - indiquer le format des balises à utiliser pour traiter les données.

2. Créer un répertoire contenant les documents au format textuel à anonymiser.
3. Lancer le script d'anonymisation au moyen de la commande suivante :
   `perl 1k_balisage.pl -r <répertoire> -e <extension des fichiers>`

Remplacement des informations identifiées

Une ou plusieurs étapes de post-traitements sont alors utiles pour procéder réellement à l'anonymisation.

1. Un script retranche à chaque date un nombre de jours aléatoirement tiré compris entre 365 et 1 460 jours (soit entre 1 et 4 ans) ; ce nombre est le même pour toutes les dates d'un document, ce qui permet de conserver les écarts temporels entre deux dates tout en observant le principe d'anonymisation. Le format des dates est reproduit à l'identique. Le script produit des fichiers d'extension « *.dat » :
   `perl 2_antidatation.pl -r <répertoire>`
2. Un second script remplace toutes les occurrences de noms et prénoms par des pseudonymes parmi l'un des 10 noms et prénoms mixtes les plus portés en France.[1] Toutes les occurrences d'un nom ou d'un prénom sont remplacées par le même pseudonyme à l'intérieur d'un document.[2] Le script produit des fichiers d'extension « *.pse » :
   `perl 4_pseudonymes.pl -r <répertoire> -e <extension des fichiers>`
3. Un dernier script remplace les données personnelles balisées par un hyperonyme (la balise typant l'information) : <*ville*>*Versailles*</*ville*> → <*ville* />. Ce script est appliqué, soit sur les fichiers d'extension « *.med » (auquel cas toutes les informations sont anonymisées), soit sur les fichiers d'extension « *.pse » (toutes les informations autres que les noms, prénoms et dates seront anonymisées, les noms, prénoms et dates ayant été préalablement traités) :
   `perl 5_hyperonymes.pl -r <répertoire> -e <extension des fichiers>`

B.4 Exemple

Le paragraphe d'exemple suivant est issu d'un compte rendu clinique en cardiologie. Toutes les informations personnelles (nom, prénom et dates) ont été modifiées par rapport à la version d'origine. Les dates qui figurent dans la version terminale anonymisée sont antidatées de 1 377 jours (soit environ 3 ans 9 mois et demi) par rapport au fichier de base ; l'écart temporel entre chaque date est néanmoins conservé.

1. Noms les plus portés : *Martin, Bernard, Dubois, Thomas, Robert, Richard, Petit, Durand, Leroy, Moreau*. Prénoms mixtes courants : *Alex, Camille, Charlie, Claude, Dominique, Louison, Maé, Maxime, Morgan, Stéphane*.

2. Dans le détail, tous les noms d'un document sont d'abord relevés puis triés par ordre alphabétique : le premier nom dans l'ordre alphabétique est remplacé par *Martin*, le second par *Bernard*, etc. Il en est de même pour les prénoms.

B.4.1 Fichier d'origine *.txt

Monsieur Théodore Bauche (21.07.53) est malheureusement revenu dans le service du 4 au 11 mai 2000 pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en octobre 99.

B.4.2 Fichier balisé *.med

Monsieur <prenom>Théodore</prenom> <nom>Bauche</nom> (<date>21.07.53</date>) est malheureusement revenu dans le service du <date>4 au 11 mai 2000</date> pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en <date>octobre 99</date>.

B.4.3 Fichier antidaté *.dat

Monsieur <prenom>Théodore</prenom> <nom>Bauche</nom> (<date>13.10.49</date>) est malheureusement revenu dans le service du <date>27 juillet au 3 août 1996</date> pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en <date>janvier 96</date>.

B.4.4 Fichier générique *.pse

Monsieur Claude Martin (<date>13.10.49</date>) est malheureusement revenu dans le service du <date>27 juillet au 3 août 1996</date> pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en <date>janvier 96</date>.

B.4.5 Fichier anonymisé *.hyp

Version post-traitée avec tous les scripts (antidatation, pseudonymes et hyperonymes)

Monsieur Claude Martin (13.10.49) est malheureusement revenu dans le service du 27 juillet au 3 août 1996 pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en janvier 96.

Version post-traitée directement après la sortie balisée (hyperonymes)

Monsieur <prenom /> <nom /> (<date />) est malheureusement revenu dans le service du <date /> pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en <date />.

B.5 Historique

18 novembre 2008 (version 1a). Création de l'outil dans le cadre du projet Akenaton pour anonymiser les comptes rendus cliniques en cardiologie.

23 novembre 2008 (version 1b). Le programme bouclait sur les expressions régulières des dates en raison des séparateurs définis dans la variable `$sep` (le point nétait pas suffisamment déspécialisé : `$sep="(\.|\-)"` au lieu de `$sep="(\\\\.|\-)"`) ; en conséquence, le point était interprété comme n'importe quel caractère et non comme un point (exemple du numéro de série *TCA020372V* dans le document 4088104749.txt).
En revanche, impossible de spécifier que le séparateur doit être le même entre plusieurs éléments d'une date : `[0-9]{2}\$sep[0-9]{2}\$3[0-9]{2,4}` renvoie comme message d'erreur *Use of uninitialized value in concatenation (.) or string*.

24 novembre 2008. Les listes utilisées ont généralement été nettoyées des mots ambigüs (càd ceux également présents dans le dictionnaire de noms communs) ; en conséquence, de véritables noms de villes (Rennes), prénoms (Sylvain) peuvent ainsi avoir été extraits de ces listes car ambigüs. En tenir compte lors de la compréhension des erreurs.
Le package `use encoding 'utf8'` permet d'indiquer à Perl que les expressions régulières contenues dans le code doivent être interprétées en UTF-8 ; il faut combiner les use open pour que les entrées/sorties soient encodées en UTF-8.

8 décembre 2008. Résolution du problème lié au mois d'août dans les expressions régulières permettant l'anonymisation des dates : lors de la récupération de la liste des mois depuis le fichier de configuration, substitution de le forme *aoÃ»t* par la forme *août* (pour rappel, le code est enregistré en UTF-8). Les packages `encoding 'utf8'`, `open ':utf8'`, `open ':std'` et `Encode 'decode_utf8'` sont restés en commentaires). Avant : *a été hospitalisé du 11 au 12 août 2004 pour complément*, après : *a été hospitalisé du <date /> pour complément*
Modification réinitialisation de la variable `@tableau=()` au lieu de `@variable=""` : évite d'avoir un enregistrement vide en début de tableau et le remplacement des espaces par une balise <hopital />

10 décembre 2008. Dans le fichier de configuration, production de deux listes de déclencheurs pour les hôpitaux : une liste longue ("Centre hospitalier") et une liste courte ("Centre") pour éviter que les déclencheurs courts prennent le dessus sur les déclencheurs longs.

12 janvier 2009. Lors de la réécriture du fichier anonymisé, supprime les espaces autour des tirets (*un traitement associant TENORMINE - ALDACTAZINE - ASPEGIC - LODALES - LEVOTHYROX.* devient *un traitement associant TENORMINE-ALDACTAZINE-ASPEGIC-LODALES-LEVOTHYROX.* dans le document 4088107098_ano.txt). Cette modification pose problème si un alignement de corpus est effectué (entre référence et résultat anonymisé) pour évaluer la qualité des résultats. Lignes commentées.

19 janvier 2009 (version 1c). Les tableaux de stockage des données ont été remplacés par des tables de hachage (`@tableau` devient `%tableau`, `@noms` devient `%noms`, etc). L'anonymisation des données par comparaison avec les références contenues dans ces tableaux s'en trouve beaucoup plus rapide (on passe de 11min 40 à seulement 3 secondes pour traiter 23 fichiers !).

11 février 2009. Améliorations ponctuelles diverses : complétion de la liste des déclencheurs de noms (Madame, Mademoiselle, Monsieur), intégration d'éléments supplémentaires lors de la seconde anonymisation, etc.

13 février 2009. Suppression de la vérification de la présence des mots dans la liste noire lorsque ces mots sont précédés d'un déclencheur (Pr, Dr, etc.) ; permet d'anonymiser *Pr Weber* alors que *Weber* figure dans la liste noire.
Ajout dans la trace du nom du fichier anonymisé pour chaque info.

21 juillet 2009. Petites retouches sur les dates qui ne se terminent pas par un séparateur mais par une fin de ligne.
La ville *Marseille* n'est pas anonymisée : ajout dans la liste des villes mais anonymisation toujours pas réalisée (pas de concordance avec la table des villes). Résolu le 23 juillet : la liste utilisée est lst_villes_sur (*Marseille* ajoutée dans cette liste).

28 juillet 2009. Lors de la récupération des noms de médicaments depuis la liste, on enregistre également la version désaccentuée du médicament (on utilise le code hexadécimal de chaque accent pour réaliser la désaccentuation : `tr/\xE8\xE9/ee/` par exemple).
Lors du test mot à mot des noms de médicaments, on teste également le mot mis en minuscules avec initiale en capitale.
Ces deux améliorations permettent de traiter efficacement le document 4088107098 dans lequel figurent des noms de médicaments en majuscules désaccentués : *ALDACTAZINE* testé sous la forme *Aldactazine* est anonymisé, de même que *ASPEGIC* testé comme *Aspegic* est trouvé comme tel dans la table de hachage des médicaments après avoir enregistré *Aspégic* sous la forme *Aspegic*.

31 juillet 2009. On crée une bijection sur les noms de médicaments composés (uniquement ceux intégrant une espace) de manière à appliquer cette bijection sur les lignes. Permet de traiter *Di Antalvic, Insuline NPH*, etc.

16 janvier 2010. Les tableaux de médicaments, prénoms et hôpitaux sont triés par tailles décroissantes des noms (même principe que dans Cokaine (COrpus and Knowledge-bAsed INformation Extraction), outil d'extraction des prescriptions médicamenteuses développé pour i2b2 2009, Deléger, Grouin, Zweigenbaum).

26 janvier 2010. Bonne gestion des passages d'arguments dans les routines (plus aucun message d'erreur).

29 janvier 2010 (version 1d). Dans le second passage, un mot commençant par une capitale suivant une balise <nom /> est remplacé par une balise <prenom /> uniquement si ce mot n'est, ni "prénom", ni "docteur" (permet d'éviter les cas : *Nom : <nom /> Prénom : <prenom />* qui devient *Nom : <nom /> <prenom /> : <prenom />* et *C. DUPONT Docteur F. DURAND* qui devient *<nom /> Docteur <nom />*).
Création de tableaux de bijection sur chaque liste de délencheurs (permet de trier par taille décroissante chaque élément).
Possibilité de trouver les expressions régulières listées en fin de ligne.

8 février 2010. Bloque sur certains fichiers, a priori en raison des parenthèses qui sont mal interprétées dans les expressions régulières.

5 mars 2010. Extension des fichiers de sortie changée de XML en SGML car pas vrai XML. Oui mais bof...

29 avril 2010. En seconde anonymisation, prise en compte des mots commençant par une capitale précédent une balise <nom /> ou <prenom />. Ces mots sont anonymisés uniquement si ils ne sont pas présents dans la liste des déclencheurs de noms.

1 septembre 2010. À partir du corpus clef en stomatologie, ajout de nouvelles entités à anonymiser (grades, numéros de dossier/acte médical) et ajout de déclencheurs supplémentaires pour les noms (métier : anesthésiste, opérateur, aide).
Les fichiers anonymisés ont pour extension « *.med » comme Medina tandis que « *.ano » est réservée comme extension de sortie de la chaîne par apprentissage (Wapiti ou CRF++).

20 décembre 2010. Modification des boucles `if` en `while` avec option "g" pour anonymiser tous les éléments pour une ligne et pas seulement le premier.
Amélioration des patrons et ajout de nouvelles règles. Permet de traiter presque tous les noms et prénoms du corpus.
Le traitement mot à mot est remonté dans la hiérarchie des opérations.

23 décembre 2010. Ajout anonymisation complémentaire sur la base de ce qui a déjà été anonymisé. permet de traiter les entités absentes des listes mais déjà traitées par des règles ou des déclencheurs (prénom *Nenci*).

10 janvier 2011. Le patron `$ligne=~/mod.le ([^\)]+)?/` pour les informations de pacemaker est trop large et anonymise des portions entières (je le commente) : *un modèle double chambre (qui permettra une stimulation de l'oreillette en cas de bradycardie sinusale liée à la majoration du traitement β-bloquant).* est anonymisé en *un modèle <info />).*

19 janvier 2011. Adaptation du programme au corpus d'anatomopathologie :
- En seconde anonymisation, on vérifie que le mot trouvé ne figure pas dans le lexique avant de le considérer comme un nom ou un prénom ;
- Le code postal doit obligatoirement être suivi par une espace et des caractères ; on ne peut pas le rencontrer en fin de ligne ou suivi par des étoiles (numéro de dossier à 5 chiffres) ;

– Les indices de grades (interne, externe) doivent commencer par une capitale pour éviter les anonymisations des adjectifs : *face, interne*. Problème également présent en stomatologie.

23 février 2012 (version 1e) et 24/02/12 (version 1f). Adaptation du script aux expériences pour le papier AMIA2012. *Les informations anonymisées sont désormais encadrées des balises typantes et non plus remplacées* comme auparavant. La précédente version remplaçant les informations par des balises, une évaluation au moyen du script de scoring nécessitait un réalignement (7e) entre le fichier « *.nom » d'origine et le fichier « *.med » anonymisé, produisant un fichier « *.enc » à évaluer. Des problèmes d'alignement ont conduit à suspendre cet alignement.

25 février 2012 (version 1g). Adaptation du script aux guidelines d'anonymisation définis pour le papier AMIA2012 et réintroduction de la seconde anonymisation.

27 février 2012 (version 1h). Ajout d'une fonction `etudePortion()` qui étudie le contexte dans lequel se trouve un patron passé en argument. Si le patron figure dans une portion déjà annotée, la fonction renvoie 1, sinon 0. Permet d'éviter d'annoter des entités à l'intérieur de portion déjà annotée ; par exemple, un prénom dans une adresse.

28 février 2012 (version 1i). Une seule règle pour les stimulateurs cardiaques : une marque de stimulateur suivie de un à cinq mots commençant par une capitale ou un chiffre et absent du dictionnaire de mots communs.
Ajout des déclencheurs *l'hôpital* et *l'Hôpital* dans le fichier de configuration (l'article est intégré à la portion annotée).

29 janvier 2012 (version 1j). Le corpus de test a été entièrement revu (1h de travail), des entités ayant été oubliées. Idem pour le corpus d'apprentissage. Modifications mineures.

12 mai 2012. Création d'un script de post-traitement : le script 2_antidatation.pl permet d'antidater toutes les dates d'un document d'un nombre de jours aléatoirement tiré entre 365 et 1460 (soit entre 1 et 4 ans). Garantit une anonymisation et une conservation des écarts temporels entre deux dates d'un document.

16 mai 2012 (version 1k). Meilleure prise en compte des dates (après application du script d'antidatation qui a révélé des erreurs dans le balisage des dates).

2 juin 2012. Création de deux scripts de post-traitements : le script 4_pseudonymes.pl remplace les occurrences de noms et prénoms par des pseudonymes, le script 5_hyperonymes.pl remplace toutes les données balisées par un hyperonyme (aucun remplacement si appliqué sur des fichiers *.dat ou *.pse).

Annexe C

Manuel d'utilisation de Medina-CRF

Sommaire

C.1 Lancement rapide

C.1.1 Préparation du tabulaire

Création du tabulaire de base (*.ali) et tokénisation du fichier de référence avec annotations (*.tok) et sans annotation (*.tkn) ; fichier.ref → fichier.ali, fichier.tok, fichier.tkn :

```
perl 10b_creeTabulaire.pl -r <répertoire>
```

Étiquetage morpho-syntaxique du fichier via Brill [1] ou le Tree Tagger [2] et création d'un tabulaire intégrant ces annotations (*.tagb) :

- Tree Tagger, fichier.ali → fichier.tagb : `perl 11a_treetagger.pl -r <répertoire>`
- Brill, fichier.tkn → fichier.tag, fichier.tagb : `perl 11b_brill.pl -r <répertoire>`

Introduction des annotations dans le tabulaire de base : fichier.ali + fichier.tagb → fichier.alig :

```
perl 12b_integreEtiquettes.pl -r <répertoire>
```

Projection des lexiques sur le tabulaire : fichier.alig → fichier.lxq :

```
perl 13_ajouteLexiques.pl -r <répertoire>
```

Production du tabulaire final *(pour l'apprentissage ou le test selon la valeur de l'option -t)* avec ajout d'informations inférées des caractéristiques de chaque token [3] : fichier.lxq → tabulaire.crf :

```
perl 14_produitTabulaire.pl -r <répertoire> -t a >tab-appr.tab
perl 14_produitTabulaire.pl -r <répertoire> -t t >tab-test.tab
```

Il est possible de créer des tabulaires pour une seule catégorie d'information *(selon la valeur renseignée pour l'option « -s »)* :

```
perl 14_produitTabulaire.pl -r <répertoire> -t a -s prenom
>tab-appr_prenom.tab
```

Afin de tester le comportement des outils d'apprentissage en fonction des tabulaires produits, deux autres options sont disponibles :

- l'option « -l » ajoute un saut de ligne entre chaque fichier traité *(génère autant de séquences que de fichiers, contre une seule séquence sans l'option)* ;
- l'option « -n » remplace les valeurs NIL dans les différentes colonnes par le token *(permet d'éviter de créer artificiellement une valeur NIL sur-représentée par rapport aux autres valeurs de la colonne).*

C.1.2 Création du modèle

Création du modèle avec CRF++ sur le tabulaire d'apprentissage *(étape assez longue si on construit un seul modèle qui englobe toutes les catégories d'information à traiter, environ 25 minutes sur les 250 fichiers de l'apprentissage)* :

```
crf_learn config.tpl tab-appr.tab modele.crf
```

Puisque plusieurs tabulaires peuvent être créés, on on intérêt à construire des modèles distincts pour chaque catégorie d'information, ce qui permet d'utiliser des configurations appropriées à chaque catégorie *(modèles adaptés et construction plus rapide, 2 minutes maximum)* :

1. Nous utilisons la version Brill réentraînée sur le corpus *Le Monde* par A. Allauzen et H. Maynard (cf. article LREC2008). L'exécutable ne fonctionne que sur une machine 32 bits.
2. Indiquer dans le fichier « config-medina.txt » l'outil d'étiquetage morpho-syntaxique à utiliser ainsi que le chemin d'accès à l'exécutable du Tree Tagger.
3. Indiquer dans le fichier « config-medina.txt » le chemin d'accès au fichier tabulaire contenant les clusters générés par l'algorithme de Brown.

```
crf_learn config_date.tpl tab-appr_date.tab modele_date.crf
crf_learn config_prenom.tpl tab-appr_prenom.tab
modele_prenom.crf
crf_learn config_ville.tpl tab-appr_ville.tab modele_ville.crf
```

C.1.3 Application du modèle

Application du modèle sur le tabulaire de test (étape rapide) :
`crf_test -m modele.crf tab-eval.tab >sortie.tab`

Si plusieurs modèles ont été construits, il faut les appliquer chacun sur le tabulaire d'évaluation. Ces différentes applications vont produire plusieurs tabulaires de sortie qu'on fusionnera par la suite.

```
crf_test -m modele_date.crf tab-eval.tab >sortie_date.tab
crf_test -m modele_prenom.crf tab-eval.tab >sortie_prenom.tab
crf_test -m modele_ville.crf tab-eval.tab >sortie_ville.tab
```

La fusion des différents tabulaires de sortie ainsi produit s'effectue au moyen d'un script (qui produit un fichier « final.tab ») :
`perl 15_fusionneTabulaires.pl`

C.1.4 Production des fichiers anonymisés

Explosion du tabulaire produit par CRF++ ; sortie.tab → fichier.ano :
`perl 16_genereFichiersAnonymises.pl -t <tabulaire>`

C.1.5 Évaluation

Concaténation des fichiers tokénisés de référence en un seul fichier, idem pour les fichiers anonymisés ; fichier.tok → all.tok, fichier.ano → all.ano :

```
cat repertoire/*ano >all.ano
cat repertoire/*tok >all.tok
```

Évaluation des deux fichiers de concaténation via l'outil de scoring Quaero [4] :
`./ne-scoring-gen anonymisation-jamia.lua <reference> <hypothese>`

C.2 Utilisation détaillée

C.2.1 Architecture globale de l'outil

L'outil se compose des éléments suivants :
- un fichier de configuration pour CRF++ : config.tpl ;
- un répertoire constitué des ressources linguistiques : data *(listes de noms d'hôpitaux, de noms de famille, de prénoms et de villes)* ;
- des scripts de création des tabulaires pour CRF++ :
  - 10b_creeTabulaire.pl : à partir des fichiers annotés de référence, crée trois fichiers ; une version tokénisée du fichier annoté de référence, une version tokénisée non annotée, et le tabulaire de base trois colonnes (offset, token, référence) ;

4. Trois arguments : le fichier de configuration (notamment les étiquettes à évaluer), la référence, et l'hypothèse.

- 11a_treeTagger.pl ou 11b_brill.pl : lance l'exécutable de Brill ou du Tree Tagger sur un répertoire et produit, pour chaque fichier du répertoire, un version étiquetée au format Brill (token/étiquette) et une version tabulaire de l'étiquetage produit ;
- 12b_integreEtiquettes.pl : intègre les annotations de Brill dans le tabulaire de base ;
- 13_ajouteLexiques.pl : vérifie la présence de chaque token en lexique *(noms, prénoms, hôpitaux, villes)* et ajoute cette information dans le précédent tabulaire ;
- 14_produitTabulaire.pl : infère des propriétés sur chaque token *(casse du token, est-un chiffre, est-une ponctuation, nombre de caractères)* et complète le tabulaire ; on obtient, soit un tabulaire pour l'apprentissage (la dernière colonne contient la réponse attendue), soit un tabulaire pour le test (cette dernière colonne est alors manquante) avec possibilité de générer des tabulaires distincts pour chaque catégorie d'information ;
- 15_fusionneTabulaires.pl : prend en entrée les différents tabulaires de sortie créés par l'application des différents modèles et génère un nouveau tabulaire « final.tab » qui fusionne, dans la dernière colonne, les réponses fournies par application de chaque modèle ; si deux modèles ont étiquetés un même token, une priorité est accordée sur certaines catégories (le prénom sur un étiquetage nom/prénom, l'hôpital sur un étiquetage hôpital/prénom ou hôpital/nom) ;
- 16_genereFichiersAnonymises.pl : à partir du tabulaire complété par CRF++ (une colonne de réponse a été ajoutée), scinde ce tabulaire en autant de fichiers qu'il y en avait à l'origine et encadre les informations des balises typantes générées par CRF++.

C.3 Exemple

Le paragraphe d'exemple suivant est issu d'un compte rendu clinique en cardiologie. Toutes les informations personnelles (nom, prénom et dates) ont été modifiées par rapport à la version d'origine.

C.3.1 Fichier de référence *.ref

Je vois aujourd'hui en consultation pré opératoire votre patient Mr <nom>Platel</nom> <prenom>Alphonsie</prenom> né le <date>2O.O4.1953</date>.

C.3.2 Fichier de référence tokénisé *.tok

Je vois aujourd'hui en consultation pré opératoire votre patient Mr <nom>Platel</nom> <prenom>Alphonsie</prenom> né le <date>2O . O4 . 1953</date> .

C.3.3 Fichier de texte tokénisé *.tkn

Je vois aujourd'hui en consultation pré opératoire votre patient Mr Platel Alphonsie né le 2O . O4 . 1953 .

C.3.4 Fichier tabulaire de base *.ali

2:0	Je	Je
2:1	vois	vois
2:2	aujourd'hui	aujourd'hui
2:3	en	en
2:4	consultation	consultation
2:5	pré	pré
2:6	opératoire	opératoire
2:7	votre	votre
2:8	patient	patient
2:9	Mr	Mr
2:10	Platel	<nom_/>
2:11	Alphonsie	<prenom_/>
2:12	né	né
2:13	le	le
2:14	2O	<date_/>
2:15	.	<date_/>
2:16	O4	<date_/>
2:17	.	<date_/>
2:18	1953	<date_/>
2:19	.	.

C.3.5 Fichier étiqueté au format Brill *.tag

Je/Pp1msn- vois/Vmip1s- aujourd'hui/Rgp en/Sp consultation/Ncfs pré/Ncms opératoire/Afpms votre/Ds2msp- patient/Afpms Mr/Ncms Platel/Npfs Alphonsie/Rgp né/Vmps-sm le/Da-ms-d 2O/Ncms ./F O4/Npms ./F 1953/Nkfp ./F

NB : l'étiquetage par le Tree Tagger ne produit pas de fichier *.tag

C.3.6 Fichier étiqueté au format tabulaire *.tagb

Étiquetage Brill

2:0	Je	Pp1msn-
2:1	vois	Vmip1s-
2:2	aujourd'hui	Rgp
2:3	en	Sp
2:4	consultation	Ncfs
2:5	pré	Ncms
2:6	opératoire	Afpms
2:7	votre	Ds2msp-
2:8	patient	Afpms
2:9	Mr	Ncms
2:10	Platel	Npfs
2:11	Alphonsie	Rgp
2:12	né	Vmps-sm
2:13	le	Da-ms-d
2:14	2O	Ncms
2:15	.	F
2:16	O4	Npms
2:17	.	F
2:18	1953	Nkfp
2:19	.	F

Étiquetage Tree Tagger

2:0	Je	PRO:PER
2:1	vois	VER:pres
2:2	aujourd'hui	ADV
2:3	en	PRP
2:4	consultation	NOM
2:5	pré	VER:pper
2:6	opératoire	NOM
2:7	votre	DET:POS
2:8	patient	ADJ
2:9	Mr	ABR
2:10	Platel	NAM
2:11	Alphonsie	NAM
2:12	né	VER:pper
2:13	le	DET:ART
2:14	2O	NOM
2:15	.	SENT
2:16	O4	NOM
2:17	.	SENT
2:18	1953	NUM
2:19	.	SENT

C.3.7 Fichier tabulaire avec annotations *.alig

Étiquetage Brill

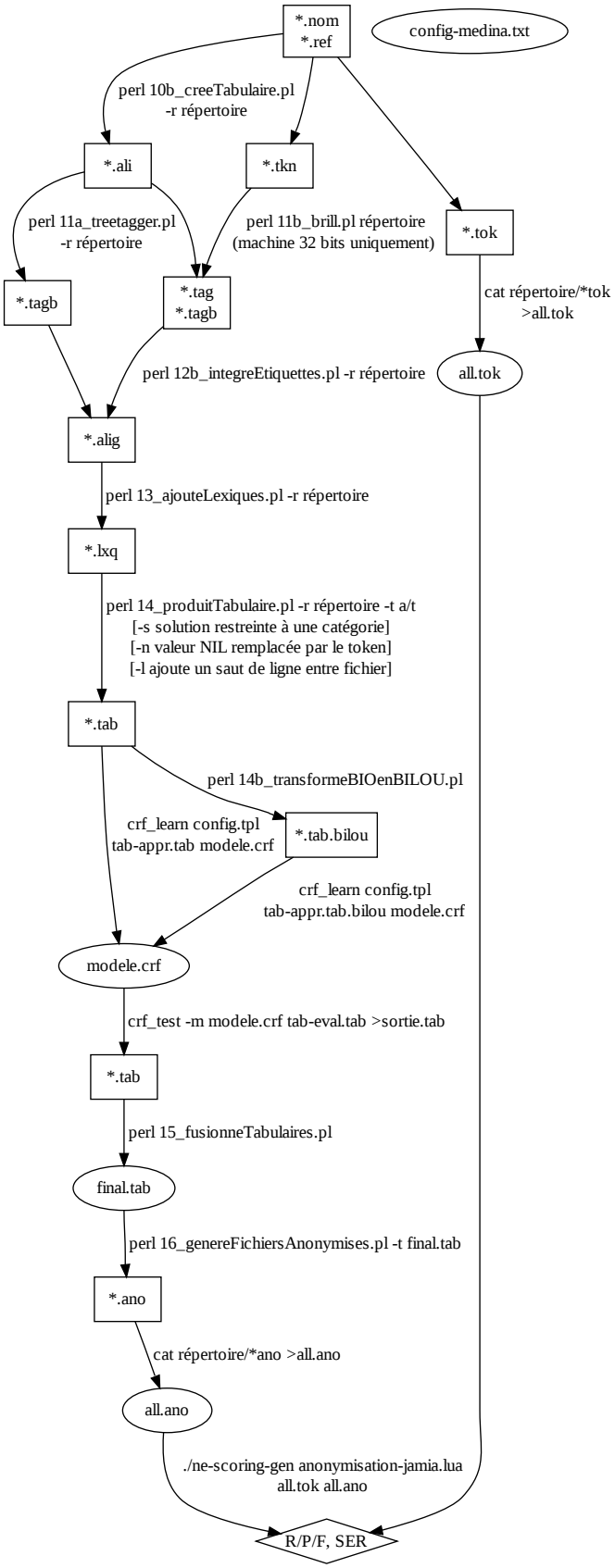
2:0	Je	Pp1msn-	Pp	Je
2:1	vois	Vmip1s-	Vm	vois
2:2	aujourd'hui	Rgp	Rg	aujourd'hui
2:3	en	Sp	Sp	en
2:4	consultation	Ncfs	Nc	consultation
2:5	pré	Ncms	Nc	pré
2:6	opératoire	Afpms	Af	opératoire
2:7	votre	Ds2msp-	Ds	votre
2:8	patient	Afpms	Af	patient
2:9	Mr	Ncms	Nc	Mr
2:10	Platel	Npfs	Np	<nom_/>
2:11	Alphonsie	Rgp	Rg	<prenom_/>
2:12	né	Vmps-sm	Vm	né
2:13	le	Da-ms-d	Da	le
2:14	2O	Ncms	Nc	<date_/>
2:15	.	F	F	<date_/>
2:16	O4	Npms	Np	<date_/>
2:17	.	F	F	<date_/>
2:18	1953	Nkfp	Nk	<date_/>
2:19	.	F	F	.

Étiquetage Tree Tagger

2:0	Je	PRO:PER	PRO	Je
2:1	vois	VER:pres	VER	vois
2:2	aujourd'hui	ADV	ADV	aujourd'hui
2:3	en	PRP	PRP	en
2:4	consultation	NOM	NOM	consultation
2:5	pré	VER:pper	VER	pré
2:6	opératoire	NOM	NOM	opératoire
2:7	votre	DET:POS	DET	votre
2:8	patient	ADJ	ADJ	patient
2:9	Mr	ABR	ABR	Mr
2:10	Platel	NAM	NAM	<nom_/>
2:11	Alphonsie	NAM	NAM	<prenom_/>
2:12	né	VER:pper	VER	né
2:13	le	DET:ART	DET	le
2:14	2O	NOM	NOM	<date_/>
2:15	.	SENT	SENT	<date_/>
2:16	O4	NOM	NOM	<date_/>
2:17	.	SENT	SENT	<date_/>
2:18	1953	NUM	NUM	<date_/>
2:19	.	SENT	SENT	.

C.3.8 Fichier tabulaire avec informations lexicales *.lxq

2:0	Je	PRO:PER	PRO	O	Je
2:1	vois	VER:pres	VER	O	vois
2:2	aujourd'hui	ADV	ADV	O	aujourd'hui
2:3	en	PRP	PRP	O	en
2:4	consultation	NOM	NOM	O	consultation
2:5	pré	VER :pper	VER	O	pré
2:6	opératoire	NOM	NOM	O	opératoire
2:7	votre	DET:POS	DET	O	votre
2:8	patient	ADJ	ADJ	O	patient
2:9	Mr	ABR	ABR	O	Mr
2:10	Platel	NAM	NAM	O	<nom_/>
2:11	Alphonsie	NAM	NAM	PRE	<prenom_/>
2:12	né	VER:pper	VER	O	né
2:13	le	DET:ART	DET	O	le
2:14	2O	NOM	NOM	O	<date_/>
2:15	.	SENT	SENT	O	<date_/>
2:16	O4	NOM	NOM	O	<date_/>
2:17	.	SENT	SENT	O	<date_/>
2:18	1953	NUM	NUM	O	<date_/>
2:19	.	SENT	SENT	O	.


*.nom
*.ref
config-medina.txt
perl 10b_creeTabulaire.pl
-r répertoire
*.ali
*.tkn
*.tok
perl 11a_treetagger.pl
-r répertoire
perl 11b_brill.pl répertoire
(machine 32 bits uniquement)
*.tagb
*.tag
*.tagb
cat répertoire/*tok
>all.tok
perl 12b_integreEtiquettes.pl -r répertoire
all.tok
*.alig
perl 13_ajouteLexiques.pl -r répertoire
*.lxq
perl 14_produitTabulaire.pl -r répertoire -t a/t
[-s solution restreinte à une catégorie]
[-n valeur NIL remplacée par le token]
[-l ajoute un saut de ligne entre fichier]
*.tab
perl 14b_transformeBIOenBILOU.pl
crf_learn config.tpl
tab-appr.tab modele.crf
*.tab.bilou
crf_learn config.tpl
tab-appr.tab.bilou modele.crf
modele.crf
crf_test -m modele.crf tab-eval.tab >sortie.tab
*.tab
perl 15_fusionneTabulaires.pl
final.tab
perl 16_genereFichiersAnonymises.pl -t final.tab
*.ano
cat répertoire/*ano >all.ano
all.ano
./ne-scoring-gen anonymisation-jamia.lua
all.tok all.ano
R/P/F, SER

C.3.9 Fichier tabulaire final *.crf

2:0	Je	PRO:PER	PRO	O	B-Mm	O	O	2	2	O	O
2:1	vois	VER:pres	VER	O	B-mm	O	O	4	2	1110110100	O
2:2	aujourd'hui	ADV	ADV	O	I-mm	O	O	11	2	1111000100	O
2:3	en	PRP	PRP	O	I-mm	O	O	2	2	1100111	O
2:4	consultation	NOM	NOM	O	I-mm	O	O	12	2	011110100	O
2:5	pré	VER:pper	VER	O	I-mm	O	O	4	2	O	O
2:6	opératoire	NOM	NOM	O	I-mm	O	O	11	2	111101100111	O
2:7	votre	DET:POS	DET	O	I-mm	O	O	5	2	011101100	O
2:8	patient	ADJ	ADJ	O	I-mm	O	O	7	2	01100110	O
2:9	Mr	ABR	ABR	O	B-Mm	O	O	2	2	O	O
2:10	Platel	NAM	NAM	O	I-Mm	O	O	6	2	O	B-nom
2:11	Alphonsie	NAM	NAM	B-PRE	I-Mm	O	O	9	2	O	B-prenom
2:12	né	VER:pper	VER	O	B-mm	O	O	3	2	110110100	O
2:13	le	DET:ART	DET	O	I-mm	O	O	2	2	101010	O
2:14	2O	NOM	NOM	O	O	O	B-DIGIT	2	2	O	B-date
2:15	.	SENT	SENT	O	O	B-PUNCT	O	1	2	1101000	I-date
2:16	O4	NOM	NOM	O	O	O	B-DIGIT	2	2	O	I-date
2:17	.	SENT	SENT	O	O	B-PUNCT	O	1	2	1101000	I-date
2:18	1953	NUM	NUM	O	O	O	B-DIGIT	4	2	O	I-date
2:19	.	SENT	SENT	O	O	B-PUNCT	O	1	2	1101000	O

C.4 Historique

2 septembre 2010. Après étude des répartitions des valeurs de chaque colonne par token, il apparaît que trois colonnes sont en distribution strictement complémentaire : 1° Appartenance du token à un lexique (HOP, MED, NOM, PRE, VIL) ; 2° Le token est une ponctuation ou pas (PUNCT) ; et 3° Le token est un chiffre arabe, romain, ou pas (DIGIT, ROMAN) :

2	HOP	NIL	NIL	-> uniquement HOP
166	MED	NIL	NIL	-> uniquement MED
1080	NIL	NIL	DIGIT	-> uniquement DIGIT
16244	NIL	NIL	NIL	-> rien
18	NIL	NIL	ROMAN	-> uniquement ROMAN
2136	NIL	PUNCT	NIL	-> uniquement PUNCT
2	NOM	NIL	NIL	-> uniquement NOM
24	PRE	NIL	NIL	-> uniquement PRE
26	VIL	NIL	NIL	-> uniquement VIL

Le contenu de ces trois colonnes est donc fusionné pour n'en former plus qu'une seule ce qui offre l'avantage de réduire le nombre de classes de features avec beaucoup de features (197 étiquettes Brill complètes, 32 étiquettes simplifiées, 25 tailles de token) par opposition aux classes avec nombre réduit de features (2 PUNCT, 3 DIGIT, 4 typo, 6 lexiques) ; on produit ainsi une colonne de 9 features. Reste donc la casse d'isolée.

19 juin 2012. Chaîne opérationnelle sur des fichiers annotés de référence dans lesquels les balises encadrent les informations à traiter.

Annexe D

Fichier de configuration CRF

D.1 Introduction

Nous reproduisons dans cette annexe le fichier de configuration utilisé par les algorithmes de CRF, dans sa version optimale, c.-à-d. la version nous ayant conduit à obtenir les meilleurs résultats au niveau global en matière d'anonymisation. Nous appelons « pivot » le token étudié au moment de l'application de chacune des règles de ce fichier de configuration.

Les coordonnées de la caractéristique utilisée par l'algorithme figure entre crochets et se décompose comme suit : [numéro de ligne,numéro de colonne] dans le tabulaire fourni en entrée.

Les bigrammes et trigrammes de caractéristiques sont représentés par une suite de coordonnées séparées par une barre oblique.

Le caractère précédent chaque numéro de règle correspond, pour le préfixe « U » à utiliser l'unigramme de la caractéristique fournie, et pour le préfixe « * » à autoriser l'algorithme à réaliser des n-grammes de caractéristiques avec cette caractéristique et les autres correspondant au token étudié.

D.2 Contenu du fichier

```
# Unigrammes de token : fenêtre de +/-3 tokens autour du pivot
U10:%x[-3,1]
U11:%x[-2,1]
*12:%x[-1,1]
*13:%x[0,1]
# Bigrammes de tokens : groupe de 2 tokens autour du pivot
U14:%x[-2,1]/%x[-1,1]
*15:%x[-1,1]/%x[0,1]
U16:%x[0,1]/%x[1,1]

# Unigrammes d'étiquettes simples : les 5 précédentes le pivot
U34:%x[0,3]
# Trigrammes d'étiquettes simples : 2 avant et après le pivot
U36:%x[-2,3]/%x[-1,3]/%x[0,3]
U37:%x[-1,3]/%x[0,3]/%x[1,3]
```

```
# Lexique : fenêtre de +/-1 token autour du pivot
U41:%x[0,4]

# Casse du token : fenêtre de +/-1 token autour du pivot
U51:%x[0,6]
# Bigrammes de casse : avant et après le pivot
U53:%x[-1,6]/%x[0,6]
U54:%x[0,6]/%x[1,6]

# Unigramme de ponctuation : avant et après le pivot
U61:%x[0,7]

# Unigramme de chiffres : fenêtre +/-1 token autour du pivot
U71:%x[0,8]

# Unigramme de taille du token : fenêtre +/-1 token autour du pivot
U81:%x[0,9]

# Unigramme de cluster
*101:%x[0,11]

# Unigramme de déclencheur
*121:%x[0,13]

# Bigram (of output)
*
```

Index

Liste des publications

Nous rassemblons dans cette section les publications réalisées pendant la thèse, classées par type de publication et par année. Les articles indexés dans la base Medline sont symbolisés par le logo PubMed.

Revues avec comités de lecture

2013

[01] Grouin C, Grabar N, Hamon T, Rosset S, Tannier X, Zweigenbaum P. Eventual situations for timeline extraction from clinical reports. In *J Am Med Inform Assoc*. 2013. Online First: April 9th. PubMed

[02] Zweigenbaum P, Lavergne T, Grabar N, Hamon T, Rosset S, Grouin C. Combining an Expert-Based Medical Entity Recognizer to a Machine-Learning System: Methods and a Case Study. In *Biomed Inform Insights*. 2013. PubMed

2012

[03] Pak A, Bernhard D, Paroubek P, Grouin C. A Combined Approach to Emotion Detection in Suicide Notes. The LIMSI participation in the i2b2/VA 2011 Challenge. In *Biomed Inform Insights*, 5(Suppl. 1):105–14. 2012. PubMed

2011

[04] Burgun A, Rosier A, Temal L, Jacques J, Messai R, Duchemin L, Deléger L, Grouin C, Van Hille P, Zweigenbaum P, Beuscart R, Delerue D, Dameron O, Mabo P, Henry C. Aide à la décision en télécardiologie par une approche basée ontologie et centrée patient. In *IRBM Ingénierie et Recherche Biomédicale*, 32(3):191–4. 2011. Elsevier-Masson.

[05] Grouin C, Zweigenbaum P. Une approche à plusieurs étapes pour anonymiser des documents médicaux. In *RSTI-RIA, Intelligence Artificielle et santé "Vers quelles applications en médecine ?"*, 25(4):525–49. 2011. Hermès-Lavoisier.

[06] Minard AL, Ligozat AL, Ben Abacha A, Bernhard D, Cartoni B, Deléger L, Grau B, Rosset S, Zweigenbaum P, Grouin C. Hybrid Methods for Improving Informa-

tion Access in Clinical Documents: Concept, Assertion, and Relation Identification. In *J Am Med Inform Assoc*, 18(5):588–93. 2011. PubMed

Actes de conférence avec comité de lecture

2013

[07] Grouin C, Zweigenbaum P. Automatic De-Identification of French Clinical Records: Comparison of Rule-Based and Machine-Learning Approaches. *Stud Health Technol Inform*. Copenhagen, Denmark. PubMed

[08] Ligozat AL, Grouin C, Garcia-Fernandez A, Bernhard D. Approches à base de fréquences pour la simplification lexicale. In *Actes TALN*, 2013. Les Sables-d'Olonne, France.

2012

[09] Deléger L, Grouin C. Detecting Negation of Medical Problems in French Clinical Notes. In *Proc of Int Health Inform symp*, 2012. Miami Beach, FL.

[10] Galibert O, Rosset S, Grouin C, Zweigenbaum P, Quintard L. Extended Named Entities Annotation on OCRed Documents: From Corpus Constitution to Evaluation Campaign. In *Proc of LREC*, 2012. Istanbul, Turkey.

[11] Ligozat AL, Grouin C, Garcia-Fernandez A, Bernhard D. ANNLOR: A Naïve Notation-system for Lexical Outputs Ranking. In *Proc of SemEval*, 2012. Montréal, Canada.

[12] Mathet Y, Widlöcher A, Fort K, François C, Galibert O, Grouin C, Kahn J, Rosset S, Zweigenbaum P. Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics. In *Proc of Coling*, 2012. Mumbai, India.

[13] Rosset S, Grouin C, Fort K, Galibert O, Kahn J, Zweigenbaum P. Structured Named Entities in two Distinct Press Corpora: Contemporary Broadcast News and Old Newspaper. In *Proc of LAW-VI*. ACL, 2012. Jeju-do, South Corea.

[14] Zweigenbaum P, Wisniewski G, Dinarelli M, Grouin C, Rosset S. Résolution des coréférences dans des comptes rendus cliniques. Une expérimentation issue du défi i2b2/VA 2011. In *Actes RFIA*, 2012. Lyon, France.

2011

[15] Galibert O, Rosset S, Grouin C, Zweigenbaum P, Quintard L. Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions. In *IJCNLP Proc*, 2011. Chiang Mai, Thailand.

[16] Grouin C, Deléger L, Cartoni B, Rosset S, Zweigenbaum P. Accès au contenu sémantique en langage de spécialité : extraction des prescriptions et concepts médicaux. In *Actes TALN*, 2011. Montpellier, France.

[17] Grouin C, Deléger L, Rosier A, Temal L, Dameron O, Van Hille P, Burgun A, Zweigenbaum P. Automatic Computation of CHA2DS2-VASc Score: Information Extraction from Clinical Texts for Thromboembolism Risk Assessment. In *AMIA Annu Symp Proc*, 2011. Washington, DC. PubMed

[18] Grouin C, Rosset S, Zweigenbaum P, Fort K, Galibert O, Quintard L. Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In *Proc of LAW-V*. ACL, 2011. Portland, OR.

Actes d'atelier avec comité de lecture

2013

[19] Grouin C. Building A Contrasting Taxa Extractor for Relation Identification from Assertions : BIOlogical Taxonomy & Ontology Phrase Extraction System. In *BioNLP Shared-Task 2013 Workshop Proc*, ACL, 2013. Sofia, Bulgaria.

2012

[20] Grouin C, Grabar N, Hamon T, Rosset S, Tannier X, Zweigenbaum P. A Tale of Temporal Relations Between Clinical Concepts and Temporal Expressions: Towards a Representation of the Clinical Patient's Timeline. In *i2b2/VA Workshop Proc*, 2012. Chicago, IL.

2011

[21] Grouin C, Dinarelli M, Rosset S, Wisniewski G, Zweigenbaum P. Coreference Resolution in Clinical Reports. The LIMSI Participation in the i2b2/VA 2011 Challenge. In *i2b2/VA Workshop Proc*, 20 octobre 2011. Washington, DC.

[22] Pak A, Bernhard D, Paroubek P, Grouin C. A Combined Approach to Emotion Detection in Suicide Notes. In *i2b2/VA Workshop Proc*, 20 octobre 2011. Washington, DC.

Actes d'atelier sans comité de lecture

2013

[23] Grouin C, Paroubek P, Zweigenbaum P. DEFT2013 passe à table : présentation du défi et des résultats. In *Actes DEFT*, 2013. Les Sables-d'Olonne, France.

Communications orales et démonstrations

2012

[24] Burgun A, Mabo P, Van Hille P, Deléger L, Grouin C, Beuscart R, Jacques J, Henry C, Duchemin L, Rosier A. *Computerized CHA2DS2Vasc Classification in Remote Atrial Fibrillation Alerts: An Ontology-Based Approach*. Congrès mondial d'électrophysiologie et de techniques cardiaques (Cardiostim), 2012. Nice, France.

[25] Grouin C. *Anonymisation automatique de documents cliniques*. Poster. Séminaire annuel de l'École doctorale « Pierre Louis de Santé Publique », Épidémiologie et Science de l'Information Biomédicale, 2012. Saint-Malo, France.

2011

[26] Dameron O, Van Hille P, Temal L, Rosier A, Deléger L, Grouin C, Zweigenbaum P, Burgun A. *Comparison of OWL and SWRL-Based Ontology Modeling Strategies for the Determination of Pacemaker Alerts Severity*. Communication orale à l'AMIA, 2011. Washington, DC.

[27] Grouin C, Deléger L, Minard AL, Ligozat AL, Ben Abacha A, Bernhard D, Cartoni B, Grau B, Rosset S, Zweigenbaum P. *Extraction d'informations médicales au LIMSI*. Démonstration industrielle TALN, 2011. Montpellier, France.

Formations suivies

Diplôme universitaire

2013

DU « *Génie Biologique et Médical* » sur la *Valorisation de la Recherche Appliquée et de l'Innovation Biomédicale*, UPMC / Faculté de Médecine Pierre et Marie Curie. Formation sous la direction du Pr Alain Sézeur. Hôpital des Diaconesses, Paris XII^{e}.

- L'innovation biomédicale ;
- Les rapports chercheurs-industries ou de l'innovation au marché ;
- Évaluation et diffusion des innovations médicales ;
- Les carrières offertes par les industries biomédicales.

Validé comme module de l'École Doctorale et comme Diplôme Universitaire *(mémoire de recherche avec soutenance)* — http://estages.ticemed.upmc.fr/gbm/

Cours en ligne

2013

"*Clinical Research Training*" course, NIH Office of Clinical Research Training and Medical Education (3 février 2013) — http://crt.nihtraining.com/

- Ethical Issues in Human Subjects Research;
- Roles and Responsibilities of the Institution;
- Roles and Responsibilities of the Investigator;
- Regulatory Issues;
- Clinical Investigators and the Mass Media.

2010

"*Protecting Human Research Participants*", National Institutes of Health (NIH) Office of Extramural Research (27 avril 2010) — http://phrp.nihtraining.com/
Certification Number: 438072.

- Codes and Regulations: The Belmont Report – Ethical Principles and Guidelines for the Protection of Human Subjects of Research; HHS Regulation for the Protection of Human Subjects, 45CFR46.
- Respect for Persons: The informed consent process; Requirements for documentation of informed consent; Waivers of informed consent; Diminished autonomy and legally authorized representatives; Participation of pregnant women in research; Assent from children and permission from parents; Obtaining informed consent from prisoners; Community consent.
- Beneficience: Risks and benefits; Privacy and Confidentiality; Institutional Review Boards (IRBs); Data and Safety Monitoring.
- Justice: Fair distribution of the benefits and burdens of research; Inclusion of Women, Minorities, and Children in Research; Issues to consider in international research.

Séminaires

2013

Epigenetic Mechanisms and Genetic Diseases. Sous la direction des Pr Jean-Louis Mandel *(chaire de génétique humaine)* et Pr Elizabeth Heard *(chaire épigénétique et mémoire cellulaire)*. Collège de France, Paris V^{e}. 21/22 mai 2013.

Séminaire en hommage à Claude Bernard. Sous la direction du Pr Alain Prochiantz *(chaire de processus morphogénétiques)*. Collège de France, Paris V^{e}. 15/16 mai 2013.

2012

Meeting semestriel Quaero. Nancy (54). 15/18 octobre 2012.

Séminaire de l'École Doctorale « Pierre Louis de Santé Publique à Paris », Saint-Malo (35). 8/10 octobre 2012.

Cyril GROUIN

Anonymisation automatique de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique

Résumé. Ce travail porte sur l'anonymisation automatique de comptes rendus cliniques. L'anonymisation consiste à masquer les informations personnelles présentes dans les documents tout en préservant les informations cliniques. Cette étape est obligatoire pour utiliser des documents cliniques en dehors du parcours de soins, qu'il s'agisse de publication de cas d'étude ou en recherche scientifique *(mise au point d'outils informatiques de traitement du contenu des dossiers, recherche de cas similaire, etc.)*. Nous avons défini douze catégories d'informations à traiter : nominatives *(noms, prénoms, etc.)* et numériques *(âges, dates, codes postaux, etc.)*. Deux approches ont été utilisées pour anonymiser les documents, l'une dite « symbolique », à base de connaissances d'expert formalisées par des expressions régulières et la projection de lexiques, l'autre par apprentissage statistique au moyen de CRF de chaîne linéaire. Plusieurs expériences ont été menées parmi lesquelles l'utilisation simple ou enchaînée de chacune des deux approches. Nous obtenons nos meilleurs résultats (F-mesure globale=0,922) en enchaînant les deux méthodes avec rassemblement des noms et prénoms en une seule catégorie (pour cette catégorie : rappel=0,953 et F-mesure=0,931). Ce travail de thèse s'accompagne de la production de plusieurs ressources : un guide d'annotation, un corpus de référence de 562 documents dont 100 annotés en double avec adjudication et calculs de taux d'accord inter-annotateurs (κ=0,807 avant fusion) et un corpus anonymisé de 17 000 comptes rendus cliniques.

Mots-clés. Anonymisation, comptes rendus médicaux, guide d'annotation, méthodes symboliques, apprentissage statistique, traitement automatique des langues.

Abstract. This work focuses on the automatic de-identification of clinical records. The de-identification consists in concealing personal information within documents while preserving clinical data. This task is mandatory so as to use clinical records outside of the patient care process, for case study publications or in scientific research *(producing automatic system to process the documents, similar cases search, etc.)*. We defined 12 categories of information to de-identify: nominative data *(last names, first names, etc.)* and numerical data *(ages, dates, zip codes, etc.)*. Two approaches have been used to de-identify the documents, an expert knowledge based method using regular expressions and lexical mapping, and a machine-learning process based upon CRF. Several experiments have been performed including the use of each approach separately or in combination. We achieved our best results (overall F-measure=0.922) while combining both approaches and merging last names and first names categories into a single one (recall=0.953 and F-measure=0.931 on this category). This work is combined with the production of several resources: a guidelines, a gold standard corpus composed of 562 documents among them 100 double annotated with adjudication and inter-annotator agreement computation (κ=0.807 before merging) and a de-identified corpus of 17,000 clinical records.

Keywords. De-identification, clinical records, guidelines, rule-based methods, machine-learning based approach, natural language processing.